Section 10-1

Correlation

PAIRED DATA

In this chapter, we will look at paired sample data (sometimes called **bivariate data**). We will address the following:

- Is there a linear relationship?
- If so, what is the equation?
- Use that equation for prediction.

CORRELATION

- A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.
- A **linear correlation** exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

EXAMPLE

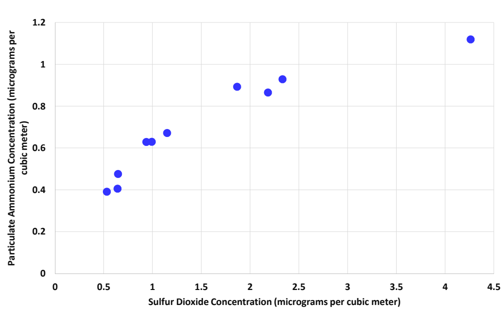
The table below gives concentration of sulfur dioxide, SO_2, (in micrograms per cubic meter) for 2008 through 2017 and concentration of particulate ammonium, NH_4, (in micrograms per cubic meter) for Georgia. Is there a correlation between the two?

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
x: SO_2	4.261	2.182	2.332	1.866	1.149	0.936	0.991	0.645	0.640	0.531
y: NH_4	1.120	0.866	0.929	0.893	0.672	0.629	0.630	0.476	0.406	0.391

SCATTERPLOT

A **scatterplot** (or **scatter diagram**) is a graph in which the paired (x, y) sample data are plotted with a horizontal x-axis and a vertical y-axis. Each individual (x, y) pair is plotted as a single point.

SCATTERPLOT OF PAIRED DATA MADE WITH EXCEL



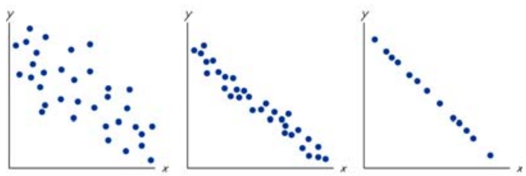
MAKING SCATTER PLOT ON THE TI-83/84

1. Select **STAT**, **1:Edit...**.
2. Enter the *x*-values for the data in **L_1** and the *y*-values in **L_2**.
3. Select **2nd**, **Y=** (for **STATPLOT**).
4. Select **Plot1**.
5. Turn Plot1 on.
6. Select the first graph **Type** which resembles a scatterplot.
7. Set **Xlist** to L_1 and **Ylist** to L_2.
8. Press **ZOOM**.
9. Select **9:ZoomStat**.

POSITIVE LINEAR CORRELATION

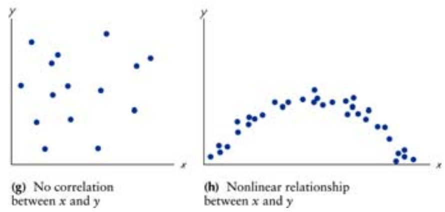
(a) Positive correlation between *x* and *y*
(b) Strong positive correlation between *x* and *y*
(c) Perfect positive correlation between *x* and *y*

NEGATIVE LINEAR CORRELATION



(d) Negative correlation between *x* and *y*
(e) Strong negative correlation between *x* and *y*
(f) Perfect negative correlation between *x* and *y*

NO LINEAR CORRELATION



LINEAR CORRELATION COEFFICIENT

The **linear correlation coefficient** r measures ***strength*** of the linear relationship between paired x and y values in a ***sample***. [The linear correlation coefficient in sometimes referred to as the **Pearson product moment correlation coefficient** in honor of Karl Pearson (1857-1936), who originally developed it.]

ASSUMPTIONS

1. The sample of paired data (x, y) is a random sample.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. Any outliers must be removed if they are known to be errors. The effects of any outliers should be considered by calculating r with and without the outliers included.

NOTATION FOR LINEAR CORRELATION COEFFICIENT

n number of pairs of data presented.

$\sum$ denotes the addition of the items indicated.

$\sum x$ denotes the sum of all x-values.

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

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

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

r represents linear correlation coefficient for a _sample_.

ρ represents linear correlation coefficient for a _population_.

LINEAR CORRELATION COEFFICIENT

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

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

The TI-83/84 calculator can compute *r*.

ρ (rho) is the linear correlation coefficient for ***all*** paired data in the ***population***.

COMPUTING THE CORRELATION COEFFICIENT *r* ON THE TI-83/84

1. Enter your x data in **L1** and your y data in **L2**.
2. Press **STAT** and arrow over to **TESTS**.
3. Select **E:LinRegTTest**.
4. Make sure that **Xlist** is set to L1, **Ylist** is set to L2, and **Freq** is set to 1.
5. Set **β & ρ** to **≠0**.
6. Leave **RegEQ** blank.
7. Arrow down to **Calculate** and press **ENTER**.
8. Press the down arrow, and you will eventually see the value for the correlation coefficient r.

ROUNDING THE LINEAR CORRELATION COEFFICIENT

- Round to three decimal places so that it can be compared to critical values in Table A-5.
- Use calculator or computer if possible.

PROPERTIES OF THE LINEAR CORRELATION COEFFICIENT

1. The value of r is always between −1 and 1 inclusive. That is, $-1 \leq r \leq 1$.
2. If all the 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 in the sense that a single outlier can dramatically affect its value.

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.

COMMON ERRORS INVOLVING CORRELATION

- **Causation**: It is wrong to conclude that correlation implies causality.
- **Averages**: Averages suppress individual variation and may inflate the correlation coefficient.
- **Linearity**: There may be ***some relationship*** between x and y even when there is no significant linear correlation.

FORMAL HYPOTHESIS TEST

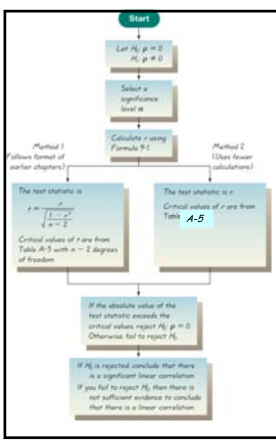
- We wish to determine whether there is a significant linear correlation between two variables.
- We present two methods.
- Both methods let H_0: $\rho = 0$

  (no significant linear correlation)

  H_1: $\rho \neq 0$

  (significant linear correlation)

TESTING FOR A LINEAR CORRELATION



METHOD 1: TEST STATISTIC IS *t*

This follows the format of Chapter 8.

Test Statistic: $t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$

Critical Values: Use Table A-3 with $n-2$ degrees of freedom.

***P*-value**: Use Table A-3 with $n-2$ degrees of freedom.

Conclusion: If $|t| >$ critical value, reject H_0 and conclude there is a linear correlation. If $|t| \leq$ critical value, fail to reject H_0; there is not sufficient evidence to conclude that there is a linear relationship.

METHOD 2: TEST STATISTIC IS *r*

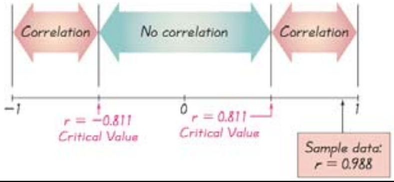
Test Statistic: r

Critical Values: Refer to Table A-5 with *no degrees of freedom*.

Conclusion: If $|r| >$ critical value, reject H_0 and conclude there is a linear correlation. If $|r| \leq$ critical value, fail to reject H_0; there is not sufficient evidence to concluded there is a linear correlation.

INTERPRETING THE LINEAR CORRELATION COEFFICIENT

- If the absolute value of r exceeds the value in Table A-5, conclude that there is a significant linear correlation.
- Otherwise, there is not sufficient evidence to support the conclusion of significant linear correlation.



CENTROID

Given a collection of paired (x, y) data, the point $(\bar{x}, \bar{y})$ is called the **centroid**.

ALTERNATIVE FORMULA FOR *r*

The formula for the correlation coefficient r can be written as

$$r = \frac{\sum\left[\frac{(x-\bar{x})}{s_x} \cdot \frac{(y-\bar{y})}{s_y}\right]}{n-1} = \frac{\sum\left(z_x \cdot z_y\right)}{n-1}$$

where s_x and s_y are the sample standard deviations of x and y, respectively; and z_x and z_y are the z scores of x and y, respectively.