We will use Formula 10-2 to help us understand the reasoning that underlies the development of the linear correlation coefficient. Because Formula 10-2 uses *z* scores, the value of $\Sigma(z_x z_y)$ does not depend on the scale that is used. Figure 10-1(a) shows the scatterplot of the shoe print and height data from Table 10-1, and Figure 10-4 shows the scatterplot of the *z* scores from the same sample data. Compare Figure 10-1(a) to Figure 10-4 and see that they are essentially the same scatterplots with different scales. The red lines in Figure 10-4 form the same coordinate axes that we have all come to know and love from earlier mathematics courses. The red lines partition Figure 10-4 into four quadrants.

If the points of the scatterplot approximate an uphill line (as in the figure), individual values of the product $z_x \cdot z_y$ tend to be positive (because most of the points are found in the first and third quadrants, where the values of z_x and z_y are either both positive or both negative), so $\Sigma(z_x z_y)$ tends to be positive. If the points of the scatterplot approximate a downhill line, most of the points are in the second and fourth quadrants, where z_x and z_y are opposite in sign, so $\Sigma(z_x z_y)$ tends to be negative. Points that follow no linear pattern tend to be scattered among the four quadrants, so the value of $\Sigma(z_x z_y)$ tends to be close to 0.

We can therefore use $\Sigma(z_x z_y)$ as a measure of how the points are configured among the four quadrants. A large positive sum suggests that the points are predominantly in the first and third quadrants (corresponding to a positive linear correlation), a large negative sum suggests that the points are predominantly in the second and fourth quadrants (corresponding to a negative linear correlation), and a sum near 0 suggests that the points are scattered among the four quadrants (with no linear correlation). We divide $\Sigma(z_x z_y)$ by $n - 1$ to get a type of average instead of a statistic that becomes larger simply because there are more data values. (The reasons for dividing by $n - 1$ instead of n are essentially the same reasons that relate to the standard deviation.) The end result is Formula 10-2, which can be algebraically manipulated into any of the other expressions for r.

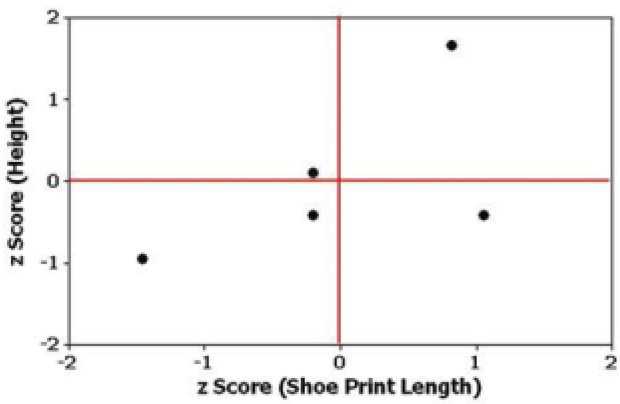

Figure 10-4 **Scatterplot of z Scores from Shoe Print Lengths and Heights in Table 10-1**

using TECHNOLOGY

STATDISK Enter the paired data in columns of the Statdisk Data Window. Select **Analysis** from the main menu bar, then use the option **Correlation and Regression.** Enter a value for the significance level. Select the columns of data to be used, then click on the **Evaluate** button. The STATDISK display will include the value of the linear correlation coefficient along with the critical value of *r*, the *P*-value, and other results to be discussed in later sections. A scatterplot can also be obtained by clicking on the **Scatterplot** button. See part of a STATDISK display in Example 1.

continued