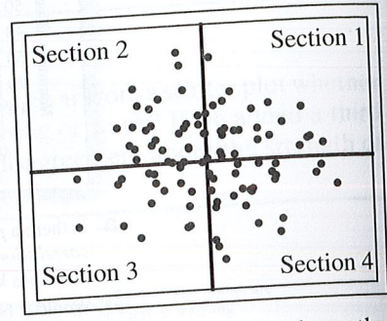
College students and middle school teachers should have a deeper understanding of correlation as something that can be measured. The next section deals with Pearson's correlation coefficient, a formula statisticians use to measure correlation between two variables. The value of this measurement indicates not only whether one variable tends to increase or decrease as the other increases, but also how strong this tendency is.

4.2 PEARSON'S CORRELATION COEFFICIENT

This section will explore the ideas behind **Pearson's correlation coefficient**, a formula used to measure the strength and direction of linear relationships between variables. Exploration 4.2 will prepare you to understand how this formula works.

Classroom Exploration 4.2

Two lines have been added to the scatter plot at the right, one vertical and the other horizontal, both passing through the center of the data points. These two lines divide the plane into four sections. In this case, the scatter plot shows no correlation between the variables. Notice that the number of data points is approximately the same for each of the four sections in the scatter plot.



1. Suppose that the data had showed a positive correlation. Which section or sections would you expect to have the most data points? Which would you expect to have the fewest? Explain why you think so.
2. Suppose that data had showed a negative correlation. Which section or sections would you expect to have the most data points? Which would you expect to have the fewest? Explain.

As we develop the formula for Pearson's correlation coefficient, you will see how the relationship you discovered in Exploration 4.2 is used to determine whether the correlation is positive or negative.

x	y
1	1
2	2
3	4
5	2
6	4
7	5

To illustrate Pearson's formula without getting too bogged down in computations, we'll use the small, highly rigged data set at the right, where most of the computations work out nicely. By the end of the section, you should be able to calculate the value of Pearson's coefficient for more complex data sets, interpret what this value indicates about the data, and have some understanding of why it works.

We'll start with what we'll call the **three S's**:

The Three S's

$$S_{xx} = \sum(x - \bar{x})^2 \qquad S_{yy} = \sum(y - \bar{y})^2 \qquad S_{xy} = \sum(x - \bar{x})(y - \bar{y})$$

You've seen S_{xx} before. In Chapter 3 that variance is given by the formula:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

S_{xx} is the numerator of this variance formula. S_{yy} is similar, but for the y-values instead of the x-values. We need both of these because we are now working with two variables instead of one (as in Chapter 3).

The third of the S's, S_{xy}, is the key to measuring correlation. To understand S_{xy}, we start with a scatter plot of the data. Just as in Exploration 4.2, we draw in two additional lines, called mean lines: a horizontal line $y = \bar{y}$, where the y-value is the average of the y-values from the data points, and a vertical line $x = \bar{x}$, where the x-value is the average of the x-values from the data points. In this case, $\bar{y} = 3$ and $\bar{x} = 4$, so the two mean lines are $y = 3$ and $x = 4$. The two mean lines divide the xy-plane (and the data) into four sections, as shown in Figure 4.2.1.

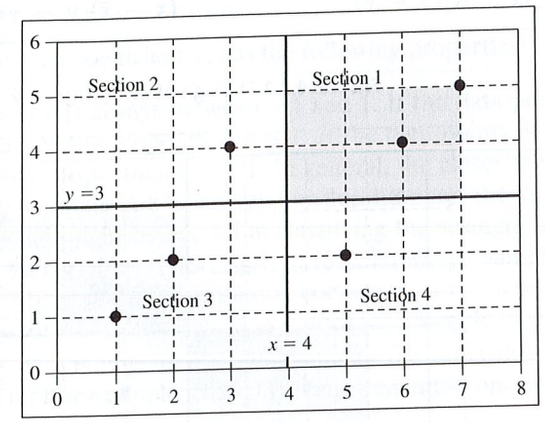


FIGURE 4.2.1 Dividing the plane into four sections

Consider the data point (7, 5) in Section 1. For this point, $x - \bar{x} = 7 - 4 = 3$, a positive number. Also, for this data point, $y - \bar{y} = 5 - 3 = 2$, also a positive number. The product, $(x - \bar{x})(y - \bar{y}) = 3 \cdot 2 = 6$, represents the area of the rectangle between the data point and the two mean lines.

For any data point (x,y) in Section 1, $x > \bar{x}$, so $(x - \bar{x})$ will be positive. Likewise, $y > \bar{y}$, so $(y - \bar{y})$ will be positive. Therefore, $(x - \bar{x})(y - \bar{y})$ will be positive. In the same way, you can determine the sign of $(x - \bar{x})(y - \bar{y})$ in each of the other sections. Fill in the blanks for Sections 2, 3, and 4 in Figure 4.2.2, just as we have already done for Section 1.

You should have concluded that $(x - \bar{x})(y - \bar{y})$ is positive for data points in Sections 1 and 3, and $(x - \bar{x})(y - \bar{y})$ is negative in Sections 2 and 4. For each data point, $(x - \bar{x})(y - \bar{y})$ represents the area of the rectangle between the data point and the two mean lines, except that in Sections 2 and 4, this area is counted as negative, as shown in Figure 4.2.3.

In this case, there are two rectangles in Section 1 with areas 6 and 2, for a total area of 8. Likewise, Section 3 has a total area of 8, while Sections 2 and 4 have a rectangle area of 1 each. The value of S_{xy} is:

$S_{xy} = \Sigma(x - \bar{x})(y - \bar{y}) = 8 + 8 + (-1) + (-1) = 14$, a positive number

Section 2 $(x - \bar{x})$ _____ $(y - \bar{y})$ _____ $(x - \bar{x})(y - \bar{y})$ _____	Section 1 $(x - \bar{x})$ positive $(y - \bar{y})$ positive $(x - \bar{x})(y - \bar{y})$ positive
Section 3 $(x - \bar{x})$ _____ $(y - \bar{y})$ _____ $(x - \bar{x})(y - \bar{y})$ _____	Section 4 $(x - \bar{x})$ _____ $(y - \bar{y})$ _____ $(x - \bar{x})(y - \bar{y})$ _____

FIGURE 4.2.2 The sign of $(x - \bar{x})(y - \bar{y})$

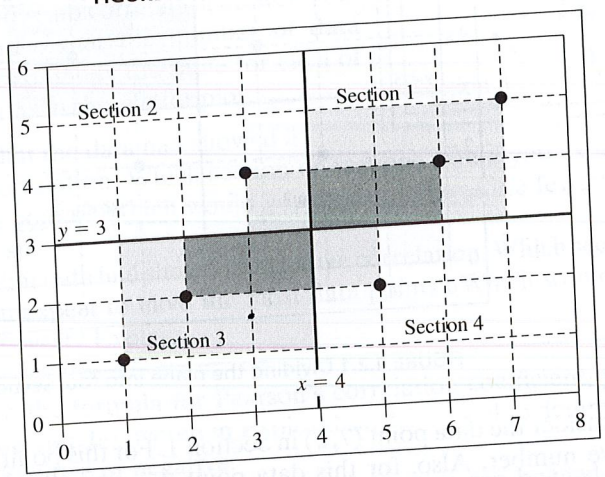


FIGURE 4.2.3 Rectangle areas represented by $(x - \bar{x})(y - \bar{y})$

As you found in Exploration 4.2, when the correlation is positive, most of the points will tend to be in Sections 1 and 3, so most of the rectangle area will be counted as positive. Whenever there is positive correlation, S_{xy} will work out to be positive. Conversely, if the correlation is negative, most of the points will tend to be in Sections 2 and 4, so most of the rectangle area will be counted as negative. Hence, S_{xy} will work out to be negative whenever there is negative correlation.

S_{xy} might be sufficient to determine whether two variables have a positive or negative correlation, but S_{xy} alone is not a good measure of the *strength* of the linear relationship between the variables. There are two reasons for this:

1. The magnitude of S_{xy} is affected by how much spread (variance) x and y have. For example, if x was a distance, we would get different values for S_{xy}, depending on whether we measured x in feet or in inches. Switching from feet to inches multiplies the value of S_{xy} by twelve. We would hesitate to say that converting x from feet to inches strengthens the linear relationship between x and y by a factor of twelve.
2. The magnitude of S_{xy} tends to increase as the number of data points increases. We wouldn't want to say that the linear relationship between two variables should be twice as strong if we collect twice as much data.

For these reasons, S_{xy} is divided by something so that the resulting value is not affected by the variances of x and y or by the number of data points. The result is known as Pearson's correlation coefficient.

Pearson's Correlation Coefficient: $$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Pearson's correlation coefficient, r, has the following properties:

1. The value of r is always between -1 and 1. If the data points lie exactly on a line with positive slope, then $r = 1$. If the data points lie exactly on a line with negative slope, then $r = -1$. In general, the closer the value of r is to 1 or -1, the stronger the linear relationship between the variables. The table in Figure 4.2.4 might be helpful in classifying the strength of the relationship between two variables (Peck, 2001, p. 157). However, sample size should also be taken into account. A correlation coefficient of $r = 0.6$ does not mean the same thing for a sample of size 3 as it does for a sample of size 300.
2. The value of r is not affected by changing the units that the variables are measured in. For example, changing measurements from feet to inches has no effect on the value of r.
3. The value of r is not affected by which variable is called x and which variable is called y.

The formulas for the three S's given previously are most useful in interpreting what they represent, but for computing the values of the three S's from data, there are some other formulas which are usually easier.

Computing Formulas for the Three S's

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$-1 < r < -0.8$	$-0.8 < r < -0.5$	$-0.5 < r < 0.5$	$0.5 < r < 0.8$	$0.8 < r < 1$
strong negative correlation	moderate negative correlation	weak or no correlation	moderate positive correlation	strong positive correlation

FIGURE 4.2.4 Values of r and strength of correlation

x	y	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$	x^2	y^2	xy
1	1	9	4	6	1	1	1
2	2	4	1	2	4	4	4
3	4	1	1	−1	9	16	12
5	2	1	1	−1	25	4	10
6	4	4	1	2	36	16	24
7	5	9	4	6	49	25	35
24	18	28	12	14	124	66	86

FIGURE 4.2.5 Computations for Pearson's correlation coefficient

Let's compute the value of Pearson's correlation coefficient both ways and compare the results. The table in Figure 4.2.5 shows the raw data and some of the computations. The bottom row gives the total for each column. Remember from earlier that $\bar{x} = 4$ and $\bar{y} = 3$.

Without going through all of the details, let's just highlight where a few of the numbers come from. The first number in the $(x - \bar{x})^2$ column uses $x = 1$ and $\bar{x} = 4$:

$$(x - \bar{x})^2 = (1 - 4)^2 = (-3)^2 = 9$$

The next number in that column uses the x-value for that row ($x = 2$):

$$(x - \bar{x})^2 = (2 - 4)^2 = (-2)^2 = 4$$

The total for that column, 28, is $\sum (x - \bar{x})^2$, which is S_{xx}. In the same way, 12, the total for the column headed $(y - \bar{y})^2$, is S_{yy}, and 14, the total for the $(x - \bar{x})(y - \bar{y})$ column, is S_{xy}.

Let's check that we get the same values by using the computing formulas for the three S's:

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 124 - \frac{24^2}{6} = 28$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 66 - \frac{18^2}{6} = 12$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 86 - \frac{(24)(18)}{6} = 14$$

As you can see, the results are the same. For those of you thinking that the original formulas are easier than the computing formulas, remember two things. First, we've done most of the computations for you; it's different when you do them all yourself. Second, this data is highly rigged so that the numbers work out nicely. Ordinarily, $\bar{x}$ and $\bar{y}$ are not nice whole numbers, but long, messy decimals. The computing formulas avoid the problem of plugging in these long, messy decimals for $\bar{x}$ and $\bar{y}$.

Finally, let's compute Pearson's coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{14}{\sqrt{28 \cdot 12}} = .764$$

This value indicates moderate (almost strong) positive correlation.

A Shortcut Using the TI-83 Plus Calculator

You can use a calculator like the TI-83 plus to get the values for $\sum x, \sum y$, etc. Hit STAT and choose 1: EDIT. Enter the x-values as L1 and the y-values as L2. Then hit STAT and then the right arrow key (to move to CALC). Notice that "2: 2-Var Stats" is on the list of options. Choose that option, and "2-Var Stats" appears on your screen. Hit ENTER and scroll down to see values for $\sum x, \sum x^2, \sum y, \sum y^2$, and $\sum xy$.

Note 1: The value S_x given by the calculator is not one of the three S's in this section, but rather the standard deviation of the x-values. We will discuss σ_x in Chapter 7.

Note 2: You'll see an even shorter shortcut in the next section.

The table in Figure 4.2.6 shows some computations based on data from *Math in Context: Statistics and the Environment, Murre Island Bats* (page 5). The first column (x) gives the air temperature in Celsius. The second column (y) shows the number of minutes that bats spend outside their caves. Some of the computations have been done for you, but there are some blank spaces left for you to fill in.

x	y	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$	x^2	y^2	xy
20	16	2.469	19.612	6.959	400	256	320
22	21	0.184	0.327	0.245	484	441	462
19	15	6.612	29.469	13.959	361	225	285
22	24	0.184	12.755	1.531	484	576	528
26	30	19.612	91.612	42.388	676	900	780
23	23	2.041	6.612	3.673	529	529	529
19	14	(a)	(b)	(c)	361	196	266
151	143	37.714	201.714	85.286	3295	3123	3170

FIGURE 4.2.6 Data and computations for Murre Island bats

Focus on Understanding

1. Using the information in Figure 4.2.6, compute the values of $\bar{x}$ and $\bar{y}$.
2. Compute the values that go in the three shaded spaces in the table.
3. Find the three S's using the original formulas.
4. Compute the three S's using the computing formulas. Check to see whether you got the same values as in #3.
5. Use the three S's to compute Pearson's correlation coefficient, r. What does this value indicate about the relationship between temperature and the time that bats spend outside their caves? Does this value seem reasonable, based on the scatter plot in Figure 4.1.1?
6. What would happen to the value of r if the temperatures were given in Fahrenheit, rather than in Celsius?
7. In this case, do you think that there is a cause-and-effect relationship between x and y? Explain.

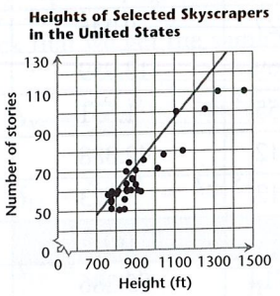
4.3 SLOPES AND EQUATIONS OF FITTED LINES

When two variables are correlated, a **fitted line** is often used to both describe the data and to predict values of one variable from the other. In middle school, the most commonly used method for fitting a line to data is similar to the way that most people hang pictures on a wall. It's not a matter of calculation, but appearance. (Does this look straight to you?) Even so, there are reasons to prefer one fitted line over another.

The exercise in Figure 4.3.1 is taken from *MathThematics, Book 2* (page 338).

12. Choose the letter of the scatter plot that you think shows the better fitted line. Explain your choice.

A. Lorna's scatter plot



B. Andrea's scatter plot

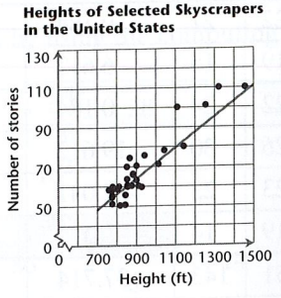


FIGURE 4.3.1 Comparing fitted lines

Classroom Exploration 4.3

1. Answer the question in Figure 4.3.1.
2. If we are going to compare fitted lines as we did in #1, it would be helpful to have some criteria for deciding whether a line fits the data well. List some general guidelines for judging by appearance (no lengthy calculations) whether a line fits the data well.
3. Does your choice for #1 fit your criteria? Explain.

Slope

Figure 4.3.2 will be helpful in understanding the concept of **slope**. If (x_1, y_1) and (x_2, y_2) are two points on a line, the change in x when moving from the first point to the second is $x_2 - x_1$. This horizontal change is also referred to as the *run*. The change in y, or *rise*, is $y_2 - y_1$. The slope m of a line is the ratio of change in y to change in x. Slope can be thought of in a number of different ways.

The Slope *m* of a Line

$$m = \frac{rise}{run} \qquad m = \frac{y_2 - y_1}{x_2 - x_1}$$

Slope is the change in y that corresponds to a 1-unit increase in x.
(To understand the last one, look at the similar triangles in Figure 4.3.2.)

For example, the points (800, 60) and (1200, 110) appear to be on Lorna's line in Figure 4.3.1. The slope of Lorna's line would be:

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{110 - 60}{1200 - 800} = \frac{50}{400} = \frac{1}{8} = 0.125$$

On Lorna's line, when x increases by 1 unit, y increases by $\frac{1}{8}$ or 0.125. Since x = *the height of a skyscraper in feet* and y = *the number of stories*, if we were using Lorna's

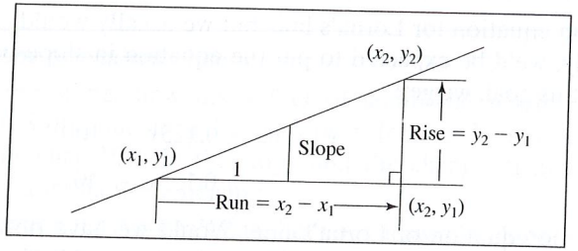


FIGURE 4.3.2 Slope of a line

line to make predictions about the number of stories in a skyscraper, we would expect an increase of 1 foot in height to go with a change of $\frac{1}{8}$ in the number of stories. In this case, it doesn't make much sense to talk about fractions of a story, so a better interpretation might be that an increase of 8 feet in height would add 1 additional story.

Ideally, we'd like to have a formula for predicting one variable from another, such as predicting the number of stories in a skyscraper from its height. To do this, we need to understand some ideas about equations of lines.

Equations of Lines

There are two forms for equations of lines that will be useful to us. They are:

The Point-Slope Equation: $y - y_1 = m(x - x_1)$

The Slope-Intercept Equation: $y = mx + b$

The **point-slope equation** is used for writing the equation of a line. In order to use this equation, we first need to know two things about the line: a point on the line and the slope of the line (hence the name of the equation). The point on the line tells us the values for x_1 and y_1 in the equation; the slope tells us the value of m.

The **slope-intercept equation** is the most useful form to have the equation in. It is set up just the way we would want to predict the value of y from the value of x. Also, by looking at the equation, we can see the slope m and the y-intercept b (hence the name slope-intercept). The ***y*-intercept** b is the value of y that goes with an x-value of zero. In an algebra class, it tells us where the line will cross the y-axis (the line $x = 0$). In statistics, we have to be a little more careful. An x-value of zero may not be reasonable in the situation. Even if it is, the line $x = 0$ may not be the same as the y-axis. If the y-axis in a scatter plot is not aligned at $x = 0$, the y-intercept will not indicate where the line meets the y-axis, but rather where it would meet the line $x = 0$.

Let's find the equation of Lorna's line. We'd like to use the point-slope equation, since it is used for exactly this purpose, writing equations for lines. Earlier, we noted that the point (800, 60) is on Lorna's line, and we found its slope to be .125. So, we have exactly the information we need to use the point-slope equation. The point (800, 60) tells us $x_1 = 800$ and $y_1 = 60$; the slope gives us $m = 0.125$. Substituting in the equation, we get:

$$y - 60 = 0.125(x - 800)$$

This is an equation for Lorna's line, but we usually would not leave it in this form. Typically, we'd be expected to put the equation in slope-intercept form. Working toward this goal, we get:

$$y - 60 = 0.125x - 100$$

$$y = 0.125x - 40$$

This is the equation of Lorna's line. Would we have obtained the same result if we had used the other point (1200, 110)? We should! You can check for yourself and see.

We could use the equation for Lorna's line to predict, for example, the number of stories in a skyscraper that was 1,000 feet tall. Since x represents the height of the skyscraper in feet, and y is the number of stories, we would substitute 1000 for x and find the value of y.

$$y = 0.125(1000) - 40 = 125 - 40 = 85$$

So, we'd predict that a 1,000-foot skyscraper would have 85 stories. Looking back on Figure 4.3.1, this seems a reasonable estimate, but it is only an estimate. Notice that, of course, the points are not exactly on Lorna's line, but are scattered both above and below the line (mostly below, in Lorna's). So we wouldn't be surprised if a 1,000-foot building had 80 or even 75 stories, but we'd probably be shocked if it had only 40. There are methods for determining how much variation from the predicted value to expect, but we will not deal with them now.

Extrapolation versus Interpolation

Suppose we used the equation for Lorna's line to predict the number of stories in a building that was 200 feet high. Working the same way as above, we get:

$$y = 0.125(200) - 40 = 25 - 40 = -15$$

Negative 15 stories! That's some weird building! What went wrong? Well, some people might blame Lorna for not drawing a better line, but the truth is that we might get a very poor prediction even from the best possible line. The reason is that all of the skyscrapers in the data set were between 700 and 1,500 feet tall. When we used the equation to make a prediction about a 1,000-foot building, we were **interpolating** within the range of heights in the data. When we try to get predictions about a 200-foot building, we are **extrapolating** well outside the range of heights in the data. In general, **interpolation** is making a prediction within the range of the data, while **extrapolation** is making a prediction outside that range. There is often no reason to expect that the same linear relationship between variables holds true outside the range of the data set. In general, extrapolation often leads to very poor predictions; interpolation is a much better bet.

Focus on Understanding

Look back at Figure 4.3.1. Andrea's line appears to go through (1000, 70) and (1500, 110).

1. Find the slope of Andrea's line.
2. Based on the slope, how many feet of additional height would be predicted for 1 additional story. Hint: Slope = (change in number of stories) ÷ (change in height). Plug in the slope and the change in number of stories (1) and solve for change in height.
3. Find an equation for Andrea's line. Put your answer in slope-intercept form.

4. What is the y-intercept of Andrea's line? Is it meaningful in this situation? Does it indicate where the line would meet the vertical axis in the scatter plot? Why or why not?
5. Use the equation to predict the number of stories in a skyscraper that is 1,200 feet tall. Do you think that a 1,200-foot building must have exactly this number of stories? Explain.
6. Use the equation to predict the number of stories in a 100-foot building. What do you think about the accuracy of this prediction and why?
7. Use the equation to predict the height of a building that has 65 stories.

Middle school students also use automatic features of calculators like the TI-83 plus to find equations for fitted lines (see *Math in Context, Insights into Data*, pages 49 and 50, for example). There are two types of lines commonly computed from data: the least squares line and the median-median line (or median fit line). We'll start with the least squares line.

4.4 THE LEAST SQUARES LINE

The fitted line most commonly used by statisticians is the **least squares line**. In some ways, it is considered to be the best possible fitted line, sometimes referred to as the line of best fit. So, what is the least squares line and what does it have to do with squares? To answer that question, we'll use some data from *MathThematics, Book 2* (page 339) as an illustration. The data are given in Figure 4.4.1.

The Idea Behind the Least Squares Line

Figure 4.4.2 consists of a scatter plot of the tent data from Figure 4.4.1 with x = *floorsize in square feet* and y = *the number of sleepers*. It also includes a fitted line that is *not* the best. The line is a bit low, especially at the right end. This scatter plot was produced using Fathom™ statistical software. Fathom has some features that are extremely helpful in explaining the idea of least squares.

Imagine this: We want to measure the vertical distance from each data point to the line. We would like to square these distances and add them up. Actually computing this sum of squares would be quite a chore, but all we really want to do right now is picture the result. Fathom does this task automatically. The results are shown in Figure 4.4.3.

Each of the squares in Figure 4.4.3 represents the squared distance between a data point and the line. The total area or sum of squares is given below the plot. The idea of "least squares" is to make this total area as small as possible.

Notice the large square at the right end. Moving the right end of the line up would make those squares smaller. Fathom allows us to drag the line around and watch what happens to the squares in the plot and to the sum of the squares (the total area of the squares) given below the plot.

14. The first two columns of the table below show the floor sizes of different tents and the number of sleepers each tent can hold.
    a. Use the data in the first two columns of the table to make a scatter plot.
    b. Draw a fitted line for your scatter plot.
    c. Using your graph, predict the floor size of a tent that can sleep 12 people.

Tent Sizes and Prices

Floor size (ft^2)	Number of sleepers	Price
135	4	$162.99
140	6	$289.99
160	8	$309.99
200	10	$299.99
108	5	$159.99
110	6	$184.99
64	4	$99.99
81	5	$139.99
100	6	$199.99
109	6	$229.99
49	3	$89.99
60	3	$109.99

15. The last column of the table above shows the price of each tent.
    a. Use the data in the last two columns to make a scatter plot.
    b. Draw a fitted line for your scatter plot.
    c. Using your graph, predict the price of a tent that can sleep 12 people.

FIGURE 4.4.1 Some data on tents

Figure 4.4.4 shows the line with the smallest sum of squares, the **least squares line**. Notice that the total area of the squares is less than the total area in the previous plot (10.13 as opposed to 16.23).

Finding the Least Squares Line

So, now that we know what the least squares line is, how would we find it without using Fathom? In Section 4.2, we introduced the three S's and their computing formulas. They are reproduced in the box on page 103 for your reference.

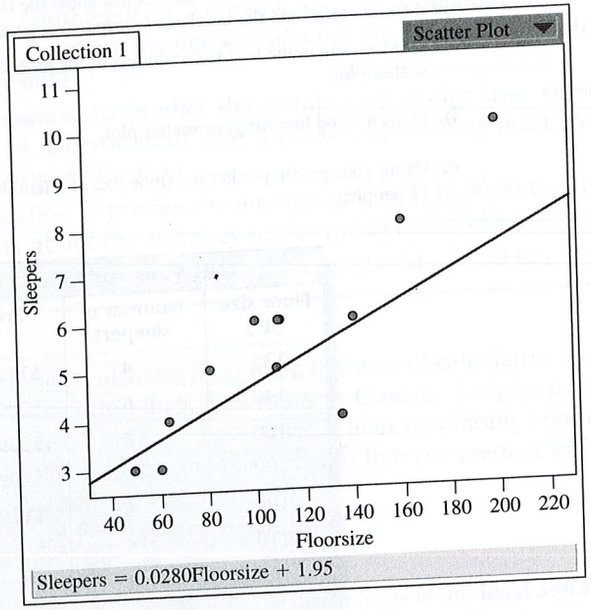


FIGURE 4.4.2 Scatter plot of the tent data

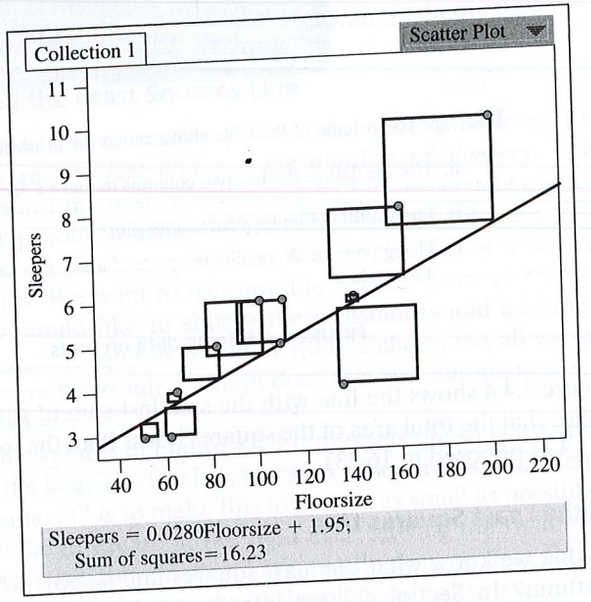


FIGURE 4.4.3 Illustrating the sum of squares

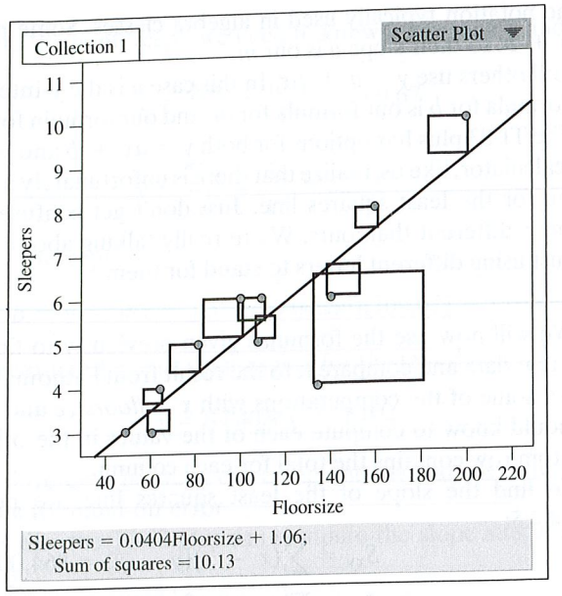


FIGURE 4.4.4 The least squares line

The Three S's and Their Computing Formulas

$$S_{xx} = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{yy} = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

Two of the three S's are used in computing the slope and y-intercept of the least squares line.

The Slope and y-intercept of the Least Squares Line: $y = mx + b$

The **slope** m is given by the formula: $m = \frac{S_{xy}}{S_{xx}}$

The ***y*-intercept** b is given by: $b = \bar{y} - m\bar{x}$

Caution

The notation for the least squares line varies from book to book and instructor to instructor. We have chosen to use the equation $y = mx + b$ to be consistent

with the notation typically used in algebra classes. Some people use the equation $y = ax + b$, so their slope a is our m.

Still others use $y = a + bx$. In this case a is the y-intercept and b is the slope! Their formula for b is our formula for m, and our formula for b is their formula for a.

The TI-83 plus has options for both $y = ax + b$ and $y = a + bx$. The makers of this calculator, like us, realize that there is unfortunately not a universally accepted notation for the least squares line. Just don't get confused if another textbook's notation is different than ours. We're really talking about exactly the same things; we're just using different letters to stand for them.

We will now use the formulas given previously to find the least squares line for the tent data and compare it to the result from Fathom. Figure 4.4.5 contains the data and some of the computations with x = *floorsize* and y = *number of sleepers*. You should know to compute each of the values in the other columns. As before, the bottom row contains the total for each column.

To find the slope of the least squares line, we need S_{xy} and S_{xx}. From Figure 4.4.5:

$$S_{xy} = \sum(x - \bar{x})(y - \bar{y}) = 864.00$$

$$S_{xx} = \sum(x - \bar{x})^2 = 21{,}406.67$$

Therefore: $m = \frac{S_{xy}}{S_{xx}} = \frac{864.00}{21{,}406.67} = 0.04036$

x	y	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$	x^2	y^2	xy
135	4	641.78	2.25	−38.00	18225	16	540
140	6	920.11	0.25	15.17	19600	36	840
160	8	2533.44	6.25	125.83	25600	64	1280
200	10	8160.11	20.25	406.50	40000	100	2000
108	5	2.78	0.25	0.83	11664	25	540
110	6	0.11	0.25	0.17	12100	36	660
64	4	2085.44	2.25	68.50	4096	16	256
81	5	821.78	0.25	14.33	6561	25	405
100	6	93.44	0.25	−4.83	10000	36	600
109	6	0.44	0.25	−0.33	11881	36	654
49	3	3680.44	6.25	151.67	2401	9	147
60	3	2466.78	6.25	124.17	3600	9	180
1316	66	21406.67	45.00	864.00	165728	408	8102

FIGURE 4.4.5 Data and computations for the least squares line

To compute the y-intercept, we need to know the means, $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{\sum x}{n} = \frac{1316}{12} = 109.67$$

$$\bar{y} = \frac{\sum y}{n} = \frac{66}{12} = 5.5$$

Therefore, the y-intercept is:

$$b = \bar{y} - m\bar{x} = 5.5 - (0.04036)(109.67) = 1.074$$

Putting things together, the least squares line for the tent data is:

$$y = 0.04036x + 1.074$$

There is a slight difference between our computations and the results from Fathom, but this is just due to round-off error.

Calculators like the TI-83 plus can compute the slope and y-intercept of this line automatically.

1. Enter the x-values as L1 and the y-values as L2.
2. Hit STAT and the right arrow key (to move to CALC). The type of calculation we want is called *linear regression*. On the TI-83 plus, this is 4: LinReg($ax + b$). Choose that option and hit ENTER. The calculator display should look something like this:

```
LinReg
y=ax+b
a=.0403612582
b=1.073715353
r²=.774936157
r=.8803045819
```

In this case, a is the slope, and b is the y-intercept. The results are consistent with what we computed earlier.

Note: If you do not see values for r^2 and r, follow the steps listed here:

1. Hit 2nd and CATALOG, the second function for the 0 key. The catalog is an alphabetical listing of calculator functions.
2. Hit the key with a green letter D above it (the x^{-1} key). This will take you to D in the alphabetical listing.
3. Scroll down until the pointer (▷) is next to the command DiagnosticOn.
4. Hit ENTER twice.

The calculator's screen should show:

```
DiagnosticOn
Done
```

Now hit STAT, move to CALC, and select 4: LinReg($ax + b$) as you did before. This time, with "diagnostics on," you should see values for r^2 and r. The diagnostics should stay on, so you should not have to turn them on again next time.

The value of r^2 has an important interpretation for statisticians, but we won't get into that in this book. The value of r should agree (except for a little round-off error) with the value of Pearson's correlation coefficient that you found in Section 4.2. This is the shorter shortcut for finding the correlation coefficient that we promised you earlier.

Focus on Understanding

1. Suppose that the table in Figure 4.4.5 did not show the columns labeled $(x - \bar{x})^2, (y - \bar{y})^2$, and $(x - \bar{x})(y - \bar{y})$. Compute S_{xy} and S_{xx} from totals in the other columns and compare your results to the values we got.
2. Use the equation of the least squares line to predict the number of sleepers for a tent with 135 square feet of floor space. Compare your answer to the first tent in the table. Do you think that this tent might comfortably fit more than four sleepers?
3. *MathThematics* asks students to predict the floorsize of a tent that sleeps 12 people. Use the equation of your least squares line to make this prediction.
4. At the end of Section 4.2, you computed $\bar{x}, \bar{y}$, and the three S's for the *Murre Island Bat* data (*Math in Context: Statistics and the Environment,* page 5). You should have found that $\bar{x} = 21.571, \bar{y} = 20.429, S_{xx} = 37.714, S_{yy} = 201.714$, and $S_{xy} = 85.286$. Recall that, in the bat data, $x =$ *air temperature in Celsius* and $y =$ *the number of minutes that bats spend outside their caves.*
   - **a.** Find the equation of the least squares line for the *Murre Island Bat* data.
   - **b.** How long would you expect bats to spend outside their caves if the air temperature was 25° C?
   - **c.** Should you use the least squares line to predict how long bats stay outside if the temperature is 100° C? Explain.
   - **d.** Enter the *Murre Island Bat* data into your calculator, and use it to find the least squares line. Compare your answer with your result from (a) above.

4.5 THE MEDIAN-MEDIAN LINE

Although the least squares line is the most common fitted line used by statisticians, it is not the only one. Another fitted line that is often presented to young students is the **median-median line**. The median-median line has several advantages:

1. It tends to satisfy all of the criteria most people would require for the *appearance* of a fitted line (provided the variables are correlated in the first place), while not involving as much computation as the least squares line.
2. As long as students know how to find medians and plot points, it is fairly easy to graph this line. This is well within the capability of middle school students. With a little more knowledge about finding slopes and equations of lines, they can find its equation.
3. Since medians are not affected by outliers, neither is the median-median line. So it may be the preferred fitted line for a data set with outliers.

Graphing the Median-Median Line

The table in Figure 4.5.1 shows the tent data we used in the last section. This time, we've sorted the data set in ascending order by x-values, and divided it into three groups: Left, Middle, and Right.

Figure 4.5.2 will clarify why we named the three groups the way we did. It shows the scatter plot with the three groups marked. In this case, there are four points in each group. It is not always possible to give every group the same number of points. In general, we try to make the groups as equal as possible. If we must have unequal groups, it's best to have more points in the low or high group than in the middle group.

In each group, we want to find a median point, a point whose x-coordinate is the median x-value and whose y-coordinate is the median y-value. For example, in the left group, the x-values arranged in order are: 49, 60, 64, 81

Group	Floorsize	Sleepers
Left	49	3
	60	3
	64	4
	81	5
Middle	100	6
	108	5
	109	6
	110	6
Right	135	4
	140	6
	160	8
	200	10

FIGURE 4.5.1 The tent data, sorted and grouped

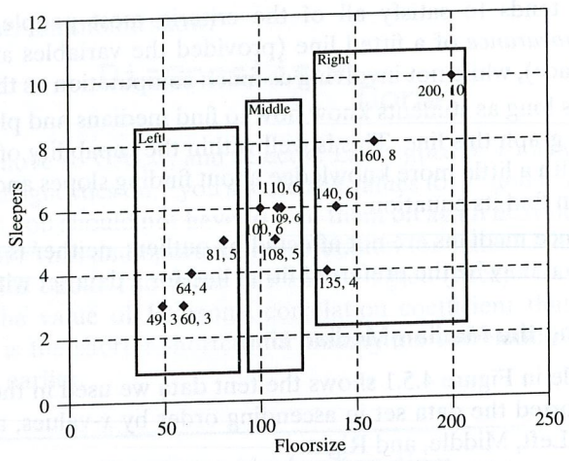


FIGURE 4.5.2 Scatter plot showing groups

Since there is not a middle value, the median is the average of the two middle ones:

$$\tilde{x} = \frac{60 + 64}{2} = 62$$

Likewise, the y-values arranged in order are:

3 3 4 5

The median y-value is:

$$\tilde{y} = \frac{3 + 4}{2} = 3.5$$

Therefore, the median point for the left group is (62, 3.5).

In the same way, we find the median points for the middle and right groups; they are (108.5, 6) and (150, 7).

Caution

The y-values in the second group are: 6 5 6 6

These must be arranged in order before finding the median: 5 6 6 6

So the median is $\tilde{y} = \frac{6+6}{2} = 6$, **not** $y = \frac{5+6}{2} = 5.5$.

Figure 4.5.3 shows these three median points, together with three lines. Line 1, the lowest of the three, is the line through the two outer median points. Line 2 is parallel to Line 1, but passes through the middle median point. Line 3 is the median-median line. It is the same as Line 1, but moved one-third of the distance in the direction of Line 2.

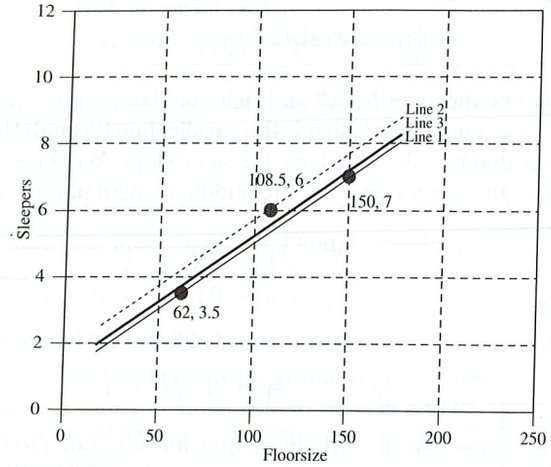


FIGURE 4.5.3 The median points and three lines

The following steps summarize the procedure for graphing the median-median line:

1. Sort the data from the smallest x-value to the largest. Use the x-values to divide the data into three groups, left, middle, and right. The groups should have as close to the same number of points as possible. If equal groups are not possible, it is better to have more points in the left group or the right group than the middle group.
2. Find the median point for each group. Remember to put the y-values in order within each group before finding the median.
3. Find the line through the two outer median points. Move this line one-third of the distance toward the middle median point. This is the median-median line.

Finding the Equation of the Median-Median Line

Continuing the previous example, we will find the equation of the median-median line. We start by finding the slope of Line 1, the line through the two outer median points. Since we know two points on this line, we can use them to find the slope:

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{7 - 3.5}{150 - 62} = .0398$$

Now, we can use the point-slope equation to write an equation for this line and put it into slope-intercept form. We'll use (150, 7) for the point on the line.

$$y - y_1 = m(x - x_1)$$
$$y - 7 = .0398(x - 150)$$

$$y - 7 = .0398x - 5.97$$
$$y = .0398x + 1.03$$

This is the line through the left and right median points.

We next want to find Line 2, the parallel line through the middle median point. Remember that parallel lines have the same slope. We'll use the point-slope formula again, using the same slope, but the middle median point (108.5, 6).

$$y - y_1 = m(x - x_1)$$
$$y - 6 = .0398(x - 108.5)$$
$$y - 6 = .0398x - 4.318$$
$$y = .0398x + 1.682$$

Finally, we know that the median-median line is parallel to the other two, but moved one-third of the distance from Line 1 toward Line 2. We can find the y-intercept for the median-median line by first finding the distance d between the y-intercepts of Lines 1 and 2:

$$d = 1.682 - 1.03 = .652$$

We want to move Line 1 up (toward Line 2) by one-third of this distance, so we add one-third of this distance to the y-intercept of Line 1.

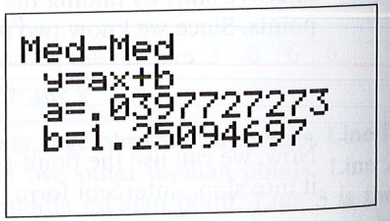
$$y = .0398x + 1.03 + \frac{1}{3}(.652)$$
$$y = .0398x + 1.247$$

This is the median-median line.

Another way to get the y-intercept for the median-median line is to do a weighted average of the y-intercepts for Line 1 and Line 2, counting Line 1 twice and Line 2 once.

$$b = \frac{1.03 + 1.03 + 1.682}{3} = 1.247 \quad \text{(the same as before)}$$

The TI-83 plus calculator can also compute the median-median line. Follow the same directions as for the least squares line, except that instead of selecting 4: LinReg, choose 3: Med-Med. You should get something like the display at the right. Just as with the least squares line, the results are slightly different because the calculator did not round off as much in its calculations as we did.



Focus on Understanding

The table at the right shows tent data from *MathThematics, Book 2* with x = *price* (to the nearest dollar) and y = *number of sleepers.*

Price	Sleepers
163	4
290	6
310	8
300	10
160	5
185	6
100	4
140	5
200	6
230	6
90	3
110	3

1. Draw a scatter plot of this data.
2. Divide the data into three groups and find the median point for each group. Mark the median points in your scatter plot.
3. Graph the median-median line in your scatter plot.
4. Use the median-median line in your graph to predict the number of people that should be able to sleep in a \$110 tent. How does the \$110 tent in the data set stack up against this prediction?
5. Based on your graph, which tent appears to be the best bargain? Explain.
6. Find the equation of the median-median line.
7. Use your equation to predict the price of a tent that sleeps ten people. Compare your answer to #5.

Chapter 4 Summary

Unlike the previous chapters, Chapter 4 explores bivariate (two variable) data sets. Scatter plots are typically used to display such data. From the distribution of the data points on the scatter plot graph, it is possible to make an intuitive judgment about the strength or weakness of the relationship (correlation) between the two variables. The strength of the relationship between the two variables can be numerically determined by using Pearson's (r) Correlation Coefficient. Quite a bit of time was spent developing the concept of where the correlation coefficient comes from and why it works the way it does. We believe that it is important for students to see where formulas come from and the ideas behind their development in order to add understanding and meaning to the complex computations. Fortunately, nowadays we let calculators and computers handle most of the routine, but long, calculations!

Ultimately, one of the outcomes of identifying a strong negative or positive relationship between two variables is to use that information to make predictions about one of the variables. Two types of predictor equations were developed in this chapter: the least squares line and the median-median line. The least squares line is the traditional statistical tool for making predictions from linear kinds of relationships. The median-median line is not as commonly used as the least-squares line in college-level elementary statistics courses, but it is introduced in some middle school mathematics courses. Median-median lines are oftentimes more intuitive for students at this level than least squares lines. The computations are also not quite as