

PART

Exploring Relationships Between Variables

Chapter 7

Scatterplots, Association, and Correlation

Chapter 8

Linear Regression

Chapter 9

Regression Wisdom

Chapter 10

Re-expressing Data: Get It Straight!

Scatterplots, Association, and Correlation



WHO Years 1970–2005

WHAT Mean error in the position of Atlantic hurricanes as predicted 72 hours ahead by the NHC

UNITS nautical miles

WHEN 1970–2005

WHERE Atlantic and Gulf of Mexico

WHY The NHC wants to improve prediction models

Look, Ma, no origin!

Scatterplots usually don't—and shouldn't—show the origin, because often neither variable has values near 0. The display should focus on the part of the coordinate plane that actually contains the data. In our example about hurricanes, none of the prediction errors or years were anywhere near 0, so the computer drew the scatterplot with axes that don't quite meet.

Hurricane Katrina killed 1,836 people¹ and caused well over 100 billion dollars in damage—the most ever recorded. Much of the damage caused by Katrina was due to its almost perfectly deadly aim at New Orleans.

Where will a hurricane go? People want to know if a hurricane is coming their way, and the National Hurricane Center (NHC) of the National Oceanic and Atmospheric Administration (NOAA) tries to predict the path a hurricane will take. But hurricanes tend to wander around aimlessly and are pushed by fronts and other weather phenomena in their area, so they are notoriously difficult to predict. Even relatively small changes in a hurricane's track can make big differences in the damage it causes.

To improve hurricane prediction, NOAA² relies on sophisticated computer models, and has been working for decades to improve them. How well are they doing? Have predictions improved in recent years? Has the improvement been consistent? Here's a timeplot of the mean error, in nautical miles, of the NHC's 72-hour predictions of Atlantic hurricanes since 1970:

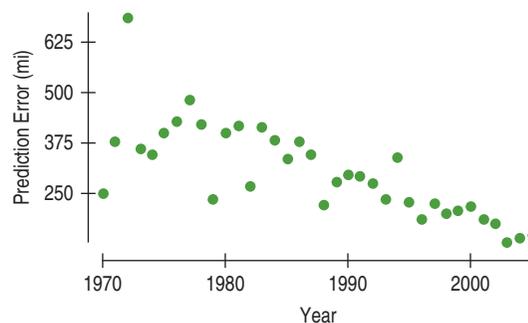


FIGURE 7.1

A scatterplot of the average error in nautical miles of the predicted position of Atlantic hurricanes for predictions made by the National Hurricane Center of NOAA, plotted against the Year in which the predictions were made.

¹ In addition, 705 are still listed as missing.

² www.nhc.noaa.gov

AS **Activity: Heights of Husbands and Wives.** Husbands are usually taller than their wives. Or are they?

Clearly, predictions have improved. The plot shows a fairly steady decline in the average error, from almost 500 nautical miles in the late 1970s to about 150 nautical miles in 2005. We can also see a few years when predictions were unusually good and that 1972 was a really bad year for predicting hurricane tracks.

This timeplot is an example of a more general kind of display called a **scatterplot**. Scatterplots may be the most common displays for data. By just looking at them, you can see patterns, trends, relationships, and even the occasional extraordinary value sitting apart from the others. As the great philosopher Yogi Berra³ once said, “You can observe a lot by watching.”⁴ Scatterplots are the best way to start observing the relationship between two *quantitative* variables.

Relationships between variables are often at the heart of what we’d like to learn from data:

- ▶ Are grades actually higher now than they used to be?
- ▶ Do people tend to reach puberty at a younger age than in previous generations?
- ▶ Does applying magnets to parts of the body relieve pain? If so, are stronger magnets more effective?
- ▶ Do students learn better with more use of computer technology?

Questions such as these relate two quantitative variables and ask whether there is an **association** between them. Scatterplots are the ideal way to *picture* such associations.

Looking at Scatterplots

AS **Activity: Making and Understanding Scatterplots.** See the best way to make scatterplots—using a computer.

Look for **Direction**: What’s my sign—positive, negative, or neither?

Look for **Form**: straight, curved, something exotic, or no pattern?

How would you describe the association of hurricane *Prediction Error* and *Year*? Everyone looks at scatterplots. But, if asked, many people would find it hard to say what to look for in a scatterplot. What do *you* see? Try to describe the scatterplot of *Prediction Error* against *Year*.

You might say that the **direction** of the association is important. Over time, the NHC’s prediction errors have decreased. A pattern like this that runs from the

upper left to the lower right  is said to be **negative**. A pattern running the other way  is called **positive**.

The second thing to look for in a scatterplot is its **form**. If there is a straight line relationship, it will appear as a cloud or swarm of points stretched out in a generally consistent, straight form. For example, the scatterplot of *Prediction Error* vs. *Year* has such an underlying **linear** form, although some points stray away from it.

Scatterplots can reveal many kinds of patterns. Often they will not be straight, but straight line patterns are both the most common and the most useful for statistics.

If the relationship isn’t straight, but curves gently, while still increasing or decreasing steadily, , we can often find ways to make it more nearly

straight. But if it curves sharply—up and then down, for example —there is much less we can say about it with the methods of this book.

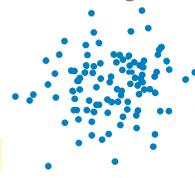
³ Hall of Fame catcher and manager of the New York Mets and Yankees.

⁴ But then he also said “I really didn’t say everything I said.” So we can’t really be sure.

Look for **Strength**: how much scatter?

The third feature to look for in a scatterplot is how strong the relationship is.

At one extreme, do the points appear tightly clustered in a single stream (whether straight, curved, or bending all over the place)? Or, at the other extreme, does the swarm of points seem to form a vague cloud through which we can



barely discern any trend or pattern?

The *Prediction error vs. Year* plot shows moderate scatter around a generally straight form. This indicates that the linear trend of improving prediction is pretty consistent and moderately strong.

Look for **Unusual Features**: Are there outliers or subgroups?

Finally, always look for the unexpected. Often the most interesting thing to see in a scatterplot is something you never thought to look for. One example of such a surprise is an **outlier** standing away from the overall pattern of the scatterplot. Such a point is almost always interesting and always deserves special attention. In the scatterplot of prediction errors, the year 1972 stands out as a year with very high prediction errors. An Internet search shows that it was a relatively quiet hurricane season. However, it included the very unusual—and deadly—Hurricane Agnes, which combined with another low-pressure center to ravage the northeastern United States, killing 122 and causing 1.3 billion 1972 dollars in damage. Possibly, Agnes was also unusually difficult to predict.

You should also look for clusters or subgroups that stand away from the rest of the plot or that show a trend in a different direction. Deviating groups should raise questions about why they are different. They may be a clue that you should split the data into subgroups instead of looking at them all together.

FOR EXAMPLE

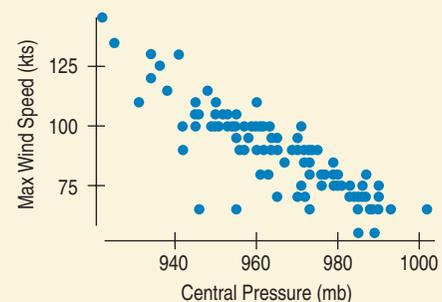
Describing the scatterplot of hurricane winds and pressure

Hurricanes develop low pressure at their centers. This pulls in moist air, pumps up their rotation, and generates high winds. Standard sea-level pressure is around 1013 millibars (mb), or 29.9 inches of mercury. Hurricane Katrina had a central pressure of 920 mb and sustained winds of 110 knots.

Here's a scatterplot of *Maximum Wind Speed (kts) vs. Central Pressure (mb)* for 163 hurricanes that have hit the United States since 1851.

Question: Describe what this plot shows.

The scatterplot shows a negative direction; in general, lower central pressure is found in hurricanes that have higher maximum wind speeds. This association is linear and moderately strong.



Roles for Variables

Which variable should go on the x -axis and which on the y -axis? What we want to know about the relationship can tell us how to make the plot. We often have questions such as:

- ▶ Do baseball teams that score more runs sell more tickets to their games?
- ▶ Do older houses sell for less than newer ones of comparable size and quality?

- ▶ Do students who score higher on their SAT tests have higher grade point averages in college?
- ▶ Can we estimate a person's percent body fat more simply by just measuring waist or wrist size?

NOTATION ALERT

So x and y are reserved letters as well, but not just for labeling the axes of a scatterplot. In Statistics, the assignment of variables to the x - and y -axes (and the choice of notation for them in formulas) often conveys information about their roles as predictor or response variable.

AS Self-Test: Scatterplot

Check. Can you identify a scatterplot's direction, form, and strength?

In these examples, the two variables play different roles. We'll call the variable of interest the **response variable** and the other the **explanatory** or **predictor variable**.⁵ We'll continue our practice of naming the variable of interest y . Naturally we'll plot it on the y -axis and place the explanatory variable on the x -axis. Sometimes, we'll call them the **x - and y -variables**. When you make a scatterplot, you can assume that those who view it will think this way, so choose which variables to assign to which axes carefully.

The roles that we choose for variables are more about how we *think* about them than about the variables themselves. Just placing a variable on the x -axis doesn't necessarily mean that it explains or predicts *anything*. And the variable on the y -axis may not respond to it in any way. We plotted prediction error on the y -axis against year on the x -axis because the National Hurricane Center is interested in how their predictions have changed over time. Could we have plotted them the other way? In this case, it's hard to imagine reversing the roles—knowing the prediction error and wanting to guess in what year it happened. But for some scatterplots, it can make sense to use either choice, so you have to think about how the choice of role helps to answer the question you have.

TI Tips

Creating a scatterplot

Let's use your calculator to make a scatterplot. First you need some data. It's okay to just enter the data in any two lists, but let's get fancy. When you are handling lots of data and several variables (as you will be soon), remembering what you stored in **L1**, **L2**, and so on can become confusing. You can—and should—give your variables meaningful names. To see how, let's store some data that you will use several times in this chapter and the next. They show the change in tuition costs at Arizona State University during the 1990s.

Naming the Lists

- Go into **STAT Edit**, place the cursor on one of the list names (**L1**, say), and use the arrow key to move to the right across all the lists until you encounter a blank column.
- Type **YR** to name this first variable, then hit **ENTER**.
- Often when we work with years it makes sense to use values like "90" (or even "0") rather than big numbers like "1990." For these data enter the years 1990 through 2000 as 0, 1, 2, . . . , 10.
- Now go to the next blank column, name this variable **TUIT**, and enter these values: 6546, 6996, 6996, 7350, 7500, 7978, 8377, 8710, 9110, 9411, 9800.

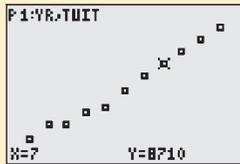
YR	TUIT	
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

Name=TUIT

YR	TUIT	
0	6546	
1	6996	
2	6996	
3	7350	
4	7500	
5	7978	
6	8377	
7		
8		
9		
10		

Name=

⁵ The x - and y -variables have sometimes been referred to as the *independent* and *dependent* variables, respectively. The idea was that the y -variable depended on the x -variable and the x -variable acted independently to make y respond. These names, however, conflict with other uses of the same terms in Statistics.



Making the Scatterplot

- Set up the **STATPLOT** by choosing the scatterplot icon (the first option).
- Identify which lists you want as **Xlist** and **Ylist**. If the data are in **L1** and **L2**, that's easy to do—but your data are stored in lists with special names. To specify your **Xlist**, go to **2nd LIST NAMES**, scroll down the list of variables until you find **YR**, then hit **ENTER**.
- Use **LIST NAMES** again to specify **Ylist:TUIT**.
- Pick a symbol for displaying the points.
- Now **ZoomStat** to see your scatterplot. (Didn't work? **ERR: DIM MISMATCH** means you don't have the same number of x 's and y 's. Go to **STAT Edit** and look carefully at your two datalists. You can easily fix the problem once you find it.)
- Notice that if you **TRACE** the scatterplot the calculator will tell you the x - and y -value at each point.

What can you Tell about the trend in tuition costs at ASU? (Remember: direction, form, and strength!)

Correlation

WHO	Students
WHAT	Height (inches), weight (pounds)
WHERE	Ithaca, NY
WHY	Data for class
HOW	Survey

Data collected from students in Statistics classes included their *Height* (in inches) and *Weight* (in pounds). It's no great surprise to discover that there is a positive association between the two. As you might suspect, taller students tend to weigh more. (If we had reversed the roles and chosen height as the explanatory variable, we might say that heavier students tend to be taller.)⁶ And the form of the scatterplot is fairly straight as well, although there seems to be a high outlier, as the plot shows.

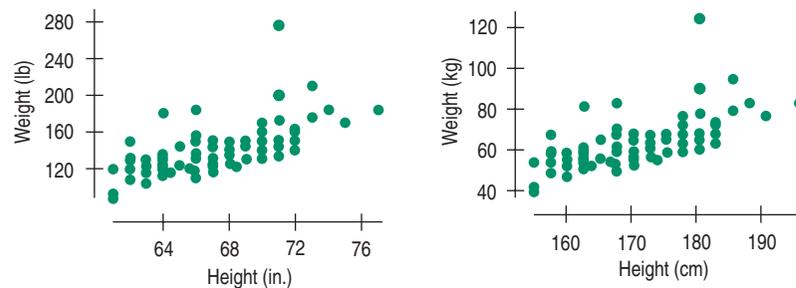


FIGURE 7.2 *Weight vs. Height of Statistics students.*

Plotting Weight vs. Height in different units doesn't change the shape of the pattern.

Activity: Correlation.

Here's a good example of how correlation works to summarize the strength of a linear relationship and disregard scaling.

The pattern in the scatterplots looks straight and is clearly a positive association, but how strong is it? If you had to put a number (say, between 0 and 1) on the strength, what would it be? Whatever measure you use shouldn't depend on the choice of units for the variables. After all, if we measure heights and weights in centimeters and kilograms instead, it doesn't change the direction, form, or strength, so it shouldn't change the number.

⁶ The son of one of the authors, when told (as he often was) that he was tall for his age, used to point out that, actually, he was young for his height.

Since the units shouldn't matter to our measure of strength, we can remove them by standardizing each variable. Now, for each point, instead of the values (x, y) we'll have the standardized coordinates (z_x, z_y) . Remember that to standardize values, we subtract the mean of each variable and then divide by its standard deviation:

$$(z_x, z_y) = \left(\frac{x - \bar{x}}{s_x}, \frac{y - \bar{y}}{s_y} \right).$$

Because standardizing makes the means of both variables 0, the center of the new scatterplot is at the origin. The scales on both axes are now standard deviation units.

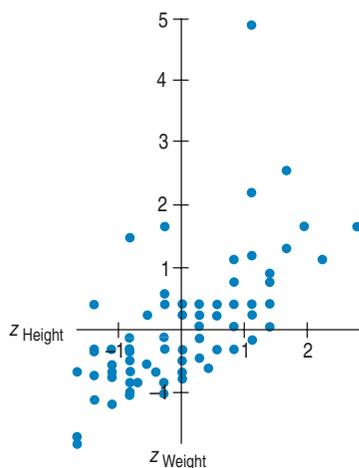
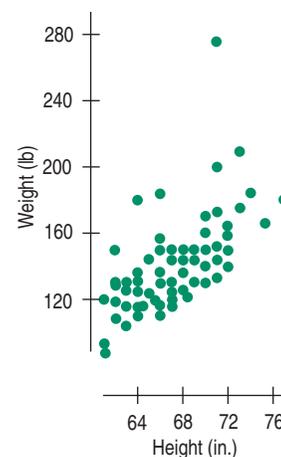


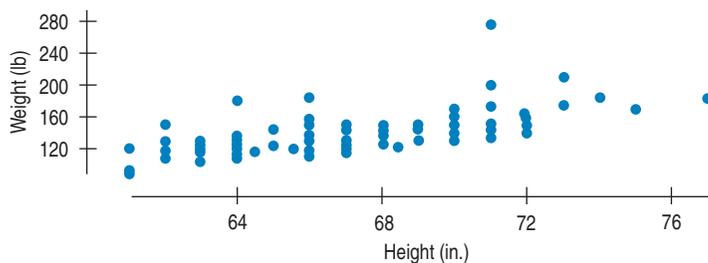
FIGURE 7.3

A scatterplot of standardized heights and weights.

Standardizing shouldn't affect the appearance of the plot. Does the plot of z-scores (Figure 7.3) look like the previous plots? Well, no. The underlying linear pattern seems steeper in the standardized plot. That's because the scales of the axes are now the same, so the length of one standard deviation is the same vertically and horizontally. When we worked in the original units, we were free to make the plot as tall and thin



or as squat and wide



as we wanted to, but that can change the impression the plot gives. By contrast,

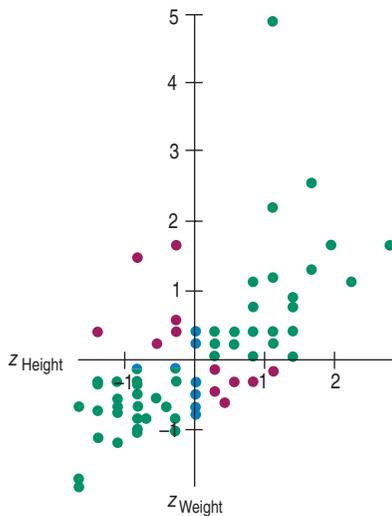


FIGURE 7.4

In this scatterplot of z-scores, points are colored according to how they affect the association: green for positive, red for negative, and blue for neutral.

AS **Activity: Correlation and Relationship Strength.** What does a correlation of 0.8 look like? How about 0.3?

NOTATION ALERT

The letter r is always used for correlation, so you can't use it for anything else in Statistics. Whenever you see an r , it's safe to assume it's a correlation.

AS **Simulation: Correlation and Linearity.** How much does straightness matter?

equal scaling gives a neutral way of drawing the scatterplot and a fairer impression of the strength of the association.⁷

Which points in the scatterplot of the z-scores give the impression of a positive association? In a positive association, y tends to increase as x increases. So, the points in the upper right and lower left (colored green) strengthen that impression. For these points, z_x and z_y have the same sign, so the product $z_x z_y$ is positive. Points far from the origin (which make the association look more positive) have bigger products.

The red points in the upper left and lower right quadrants tend to weaken the positive association (or support a negative association). For these points, z_x and z_y have opposite signs. So the product $z_x z_y$ for these points is negative. Points far from the origin (which make the association look more negative) have a negative product even larger in magnitude.

Points with z-scores of zero on either variable don't vote either way, because $z_x z_y = 0$. They're colored blue.

To turn these products into a measure of the strength of the association, just add up the $z_x z_y$ products for every point in the scatterplot:

$$\sum z_x z_y.$$

This summarizes the direction *and* strength of the association for all the points. If most of the points are in the green quadrants, the sum will tend to be positive. If most are in the red quadrants, it will tend to be negative.

But the *size* of this sum gets bigger the more data we have. To adjust for this, the natural (for statisticians anyway) thing to do is to divide the sum by $n - 1$.⁸ The ratio is the famous **correlation coefficient**:

$$r = \frac{\sum z_x z_y}{n - 1}.$$

For the students' heights and weights, the correlation is 0.644. There are a number of alternative formulas for the correlation coefficient, but this form using z-scores is best for understanding what correlation means.

Correlation Conditions

Correlation measures the strength of the *linear* association between two *quantitative* variables. Before you use correlation, you must check several *conditions*:

- ▶ **Quantitative Variables Condition:** Are both variables quantitative? Correlation applies only to quantitative variables. Don't apply correlation to categorical data masquerading as quantitative. Check that you know the variables' units and what they measure.
- ▶ **Straight Enough Condition:** Is the form of the scatterplot straight enough that a linear relationship makes sense? Sure, you can *calculate* a correlation coefficient for any pair of variables. But correlation measures the strength only

⁷ When we draw a scatterplot, what often looks best is to make the length of the x -axis slightly larger than the length of the y -axis. This is an aesthetic choice, probably related to the Golden Ratio of the Greeks.

⁸ Yes, the same $n - 1$ as in the standard deviation calculation. And we offer the same promise to explain it later.

AS

Case Study: Mortality and Education. Is the mortality rate lower in cities with higher education levels?

of the *linear* association, and will be misleading if the relationship is not linear. What is “straight enough”? How non-straight would the scatterplot have to be to fail the condition? This is a judgment call that you just have to think about. Do you think that the underlying relationship is curved? If so, then summarizing its strength with a correlation would be misleading.

- ▶ **Outlier Condition:** Outliers can distort the correlation dramatically. An outlier can make an otherwise weak correlation look big or hide a strong correlation. It can even give an otherwise positive association a negative correlation coefficient (and vice versa). When you see an outlier, it’s often a good idea to report the correlation with and without that point.

Each of these conditions is easy to check with a scatterplot. Many correlations are reported without supporting data or plots. Nevertheless, you should still think about the conditions. And you should be cautious in interpreting (or accepting others’ interpretations of) the correlation when you can’t check the conditions for yourself.

FOR EXAMPLE

Correlating wind speed and pressure

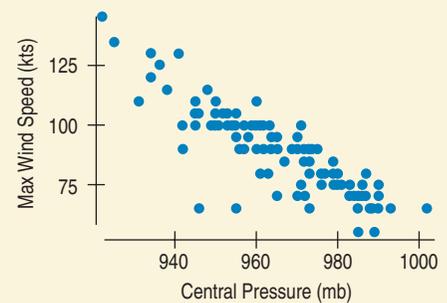
Recap: We looked at the scatterplot displaying hurricane wind speeds and central pressures.

The correlation coefficient for these wind speeds and pressures is $r = -0.879$.

Question: Check the conditions for using correlation. If you feel they are satisfied, interpret this correlation.

- ▶ **Quantitative Variables Condition:** Both wind speed and central pressure are quantitative variables, measured (respectively) in knots and millibars.
- ▶ **Straight Enough Condition:** The pattern in the scatterplot is quite straight.
- ▶ **Outlier Condition:** A few hurricanes seem to straggle away from the main pattern, but they don’t appear to be extreme enough to be called outliers. It may be worthwhile to check on them, however.

The conditions for using correlation are satisfied. The correlation coefficient of $r = -0.879$ indicates quite a strong negative linear association between the wind speeds of hurricanes and their central pressures.



JUST CHECKING

Your Statistics teacher tells you that the correlation between the scores (points out of 50) on Exam 1 and Exam 2 was 0.75.

1. Before answering any questions about the correlation, what would you like to see? Why?
2. If she adds 10 points to each Exam 1 score, how will this change the correlation?
3. If she standardizes scores on each exam, how will this affect the correlation?
4. In general, if someone did poorly on Exam 1, are they likely to have done poorly or well on Exam 2? Explain.
5. If someone did poorly on Exam 1, can you be sure that they did poorly on Exam 2 as well? Explain.

STEP-BY-STEP EXAMPLE

Looking at Association

When your blood pressure is measured, it is reported as two values: systolic blood pressure and diastolic blood pressure.

Questions: How are these variables related to each other? Do they tend to be both high or both low? How strongly associated are they?

THINK

Plan State what you are trying to investigate.

Variables Identify the two quantitative variables whose relationship we wish to examine. Report the *W*'s, and be sure both variables are recorded for the same individuals.

Plot Make the scatterplot. Use a computer program or graphing calculator if you can.

Check the conditions.

REALITY CHECK

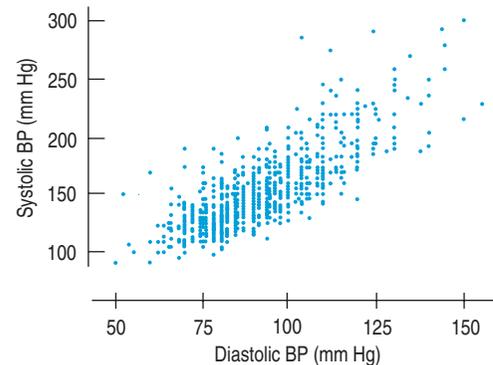
Looks like a strong positive linear association. We shouldn't be surprised if the correlation coefficient is positive and fairly large.

SHOW

Mechanics We usually calculate correlations with technology. Here we have 1406 cases, so we'd never try it by hand.

I'll examine the relationship between two measures of blood pressure.

The variables are systolic and diastolic blood pressure (SBP and DBP), recorded in millimeters of mercury (mm Hg) for each of 1406 participants in the Framingham Heart Study, a famous health study in Framingham, MA.⁹



- ✓ **Quantitative Variables Condition:** Both SBP and DBP are quantitative and measured in mm Hg.
- ✓ **Straight Enough Condition:** The scatterplot looks straight.
- ✓ **Outlier Condition:** There are a few straggling points, but none far enough from the body of the data to be called outliers.

I have two quantitative variables that satisfy the conditions, so correlation is a suitable measure of association.

The correlation coefficient is $r = 0.792$.

⁹ www.nhlbi.nih.gov/about/framingham



Conclusion Describe the direction, form, and strength you see in the plot, along with any unusual points or features. Be sure to state your interpretations in the proper context.

The scatterplot shows a positive direction, with higher *SBP* going with higher *DBP*. The plot is generally straight, with a moderate amount of scatter. The correlation of 0.792 is consistent with what I saw in the scatterplot. A few cases stand out with unusually high *SBP* compared with their *DBP*. It seems far less common for the *DBP* to be high by itself.

TI Tips

Finding the correlation

```
CATALOG
DependAuto
det(
DiagnosticOff
DiagnosticOn
dim(
Disp
DispGraph
```

```
DiagnosticOn Done
█
```



```
EDIT 0:CALC TESTS
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8:LinReg(a+bx)
9:LnReg
0:ExpReg
```

```
LinReg(a+bx) L1R
, L2UIT █
```

```
LinReg
y=a+bx
a=6439.954545
b=326.0818182
r^2=.9863642357
r=.9931587163
█
```

Now let's use the calculator to find a correlation. Unfortunately, the statistics package on your TI calculator does not automatically do that. Correlations are one of the most important things we might want to do, so here's how to fix that, once and for all.

- Hit **2nd CATALOG** (on the zero key). You now see a list of everything the calculator knows how to do. Impressive, huh?
- Scroll down until you find **DiagnosticOn**. Hit **ENTER**. Again. It should say **Done**.

Now and forevermore (or perhaps until you change batteries) your calculator will find correlations.

Finding the Correlation

- *Always* check the conditions first. Look at the scatterplot for the Arizona State tuition data again. Does this association look linear? Are there outliers? This plot looks fine, but remember that correlation can be used to describe the strength of *linear* associations only, and outliers can distort the results. Eyeballing the scatterplot is an essential first step. (You should be getting used to checking on assumptions and conditions before jumping into a statistical procedure—it's always important.)
- Under the **STAT CALC** menu, select **8:LinReg(a+bx)** and hit **ENTER**.
- Now specify x and y by importing the names of your variables from the **LIST NAMES** menu. First name your x -variable followed by a comma, then your y -variable, creating the command

LinReg(a+bx) L1R, L2UIT

Wow! A lot of stuff happened. If you suspect all those other numbers are important, too, you'll really enjoy the next chapter. But for now, it's the value of r you care about. What does this correlation, $r = 0.993$, say about the trend in tuition costs?

Correlation Properties

AS **Activity: Construct Scatterplots with a Given Correlation.** Try to make a scatterplot that has a given correlation. How close can you get?

Height and Weight, Again

We could have measured the students' weights in stones. In the now outdated UK system of measures, a stone is a measure equal to 14 pounds. And we could have measured heights in hands. Hands are still commonly used to measure the heights of horses. A hand is 4 inches. But no matter what *units* we use to measure the two variables, the *correlation* stays the same.

TI-*inspire*

Correlation and Scatterplots. See how the correlation changes as you drag data points around in a scatterplot.

Here's a useful list of facts about the correlation coefficient:

- ▶ The sign of a correlation coefficient gives the direction of the association.
- ▶ Correlation is always between -1 and $+1$. Correlation *can* be exactly equal to -1.0 or $+1.0$, but these values are unusual in real data because they mean that all the data points fall *exactly* on a single straight line.
- ▶ Correlation treats x and y symmetrically. The correlation of x with y is the same as the correlation of y with x .
- ▶ Correlation has no units. This fact can be especially appropriate when the data's units are somewhat vague to begin with (IQ score, personality index, socialization, and so on). Correlation is sometimes given as a percentage, but you probably shouldn't do that because it suggests a percentage of *something*—and correlation, lacking units, has no “something” of which to be a percentage.
- ▶ Correlation is not affected by changes in the center or scale of either variable. Changing the units or baseline of either variable has no effect on the correlation coefficient. Correlation depends only on the z -scores, and they are unaffected by changes in center or scale.
- ▶ Correlation measures the strength of the *linear* association between the two variables. Variables can be strongly associated but still have a small correlation if the association isn't linear.
- ▶ Correlation is sensitive to outliers. A single outlying value can make a small correlation large or make a large one small.

How strong is strong? You'll often see correlations characterized as “weak,” “moderate,” or “strong,” but be careful. There's no agreement on what those terms mean. The same numerical correlation might be strong in one context and weak in another. You might be thrilled to discover a correlation of 0.7 between the new summary of the economy you've come up with and stock market prices, but you'd consider it a design failure if you found a correlation of “only” 0.7 between two tests intended to measure the same skill. Deliberately vague terms like “weak,” “moderate,” or “strong” that describe a linear association can be useful additions to the numerical summary that correlation provides. But be sure to include the correlation and show a scatterplot, so others can judge for themselves.

FOR EXAMPLE

Changing scales

Recap: We found a correlation of $r = -0.879$ between hurricane wind speeds in knots and their central pressures in millibars.

Question: Suppose we wanted to consider the wind speeds in miles per hour (1 mile per hour = 0.869 knots) and central pressures in inches of mercury (1 inch of mercury = 33.86 millibars). How would that conversion affect the conditions, the value of r , and our interpretation of the correlation coefficient?

Not at all! Correlation is based on standardized values (z -scores), so the conditions, the value of r , and the proper interpretation are all unaffected by changes in units.

Warning: Correlation \neq Causation

Whenever we have a strong correlation, it's tempting to try to explain it by imagining that the predictor variable has *caused* the response to change. Humans are like that; we tend to see causes and effects in everything.

Sometimes this tendency can be amusing. A scatterplot of the human population (y) of Oldenburg, Germany, in the beginning of the 1930s plotted against the number of storks nesting in the town (x) shows a tempting pattern.

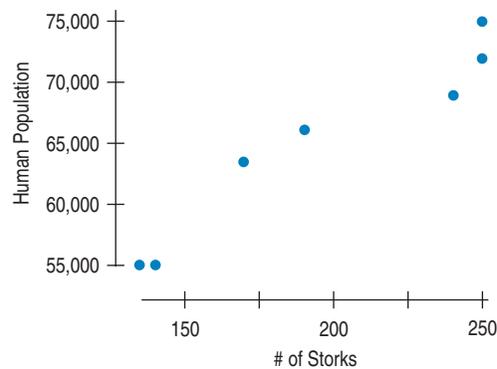


FIGURE 7.5

The number of storks in Oldenburg, Germany, plotted against the population of the town for 7 years in the 1930s. The association is clear. How about the causation? (Ornithologishe Monatsberichte, 44, no. 2)

Anyone who has seen the beginning of the movie *Dumbo* remembers Mrs. Jumbo anxiously waiting for the stork to bring her new baby. Even though you know it's silly, you can't help but think for a minute that this plot shows that storks are the culprits. The two variables are obviously related to each other (the correlation is 0.97!), but that doesn't prove that storks bring babies.

It turns out that storks nest on house chimneys. More people means more houses, more nesting sites, and so more storks. The causation is actually in the *opposite* direction, but you can't tell from the scatterplot or correlation. You need additional information—not just the data—to determine the real mechanism.

A scatterplot of the damage (in dollars) caused to a house by fire would show a strong correlation with the number of firefighters at the scene. Surely the damage doesn't cause firefighters. And firefighters do seem to cause damage, spraying water all around and chopping holes. Does that mean we shouldn't call the fire department? Of course not. There is an underlying variable that leads to both more damage and more firefighters: the size of the blaze.

A hidden variable that stands behind a relationship and determines it by simultaneously affecting the other two variables is called a **lurking variable**. You can often debunk claims made about data by finding a lurking variable behind the scenes.

Scatterplots and correlation coefficients *never* prove causation. That's one reason it took so long for the U.S. Surgeon General to get warning labels on cigarettes. Although there was plenty of evidence that increased smoking was *associated* with increased levels of lung cancer, it took years to provide evidence that smoking actually *causes* lung cancer.

Does cancer cause smoking? Even if the correlation of two variables is due to a causal relationship, the correlation itself cannot tell us what causes what.

Sir Ronald Aylmer Fisher (1890–1962) was one of the greatest statisticians of the 20th century. Fisher testified in court (in testimony paid for by the tobacco companies) that a causal relationship might underlie the correlation of smoking and cancer:

“Is it possible, then, that lung cancer . . . is one of the causes of smoking cigarettes? I don't think it can be excluded . . . the pre-cancerous condition is one involving a certain amount of slight chronic inflammation . . .

A slight cause of irritation . . . is commonly accompanied by pulling out a cigarette, and getting a little compensation for life's minor ills in that way. And . . . is not unlikely to be associated with smoking more frequently."

Ironically, the proof that smoking indeed is the cause of many cancers came from experiments conducted following the principles of experiment design and analysis that Fisher himself developed—and that we'll see in Chapter 13.

Correlation Tables

It is common in some fields to compute the correlations between every pair of variables in a collection of variables and arrange these correlations in a table. The rows and columns of the table name the variables, and the cells hold the correlations.

Correlation tables are compact and give a lot of summary information at a glance. They can be an efficient way to start to look at a large data set, but a dangerous one. By presenting all of these correlations without any checks for linearity and outliers, the correlation table risks showing truly small correlations that have been inflated by outliers, truly large correlations that are hidden by outliers, and correlations of any size that may be meaningless because the underlying form is not linear.

	Assets	Sales	Market Value	Profits	Cash Flow	Employees
Assets	1.000					
Sales	0.746	1.000				
Market Value	0.682	0.879	1.000			
Profits	0.602	0.814	0.968	1.000		
Cash Flow	0.641	0.855	0.970	0.989	1.000	
Employees	0.594	0.924	0.818	0.762	0.787	1.000

Table 7.1

A correlation table of data reported by *Forbes* magazine for large companies. From this table, can you be sure that the variables are linearly associated and free from outliers?

The diagonal cells of a correlation table always show correlations of exactly 1. (Can you see why?) Correlation tables are commonly offered by statistics packages on computers. These same packages often offer simple ways to make all the scatterplots that go with these correlations.

Straightening Scatterplots

Correlation is a suitable measure of strength for straight relationships only. When a scatterplot shows a bent form that consistently increases or decreases, we can often straighten the form of the plot by re-expressing one or both variables.

Some camera lenses have an adjustable aperture, the hole that lets the light in. The size of the aperture is expressed in a mysterious number called the *f/stop*. Each increase of one *f/stop* number corresponds to a halving of the light that is allowed to come through. The *f/stops* of one digital camera are

f/stop: 2.8 4 5.6 8 11 16 22 32

When you halve the shutter speed, you cut down the light, so you have to open the aperture one notch. We could experiment to find the best f /stop value for each shutter speed. A table of recommended shutter speeds and f /stops for a camera lists the relationship like this:

Shutter speed:	1/1000	1/500	1/250	1/125	1/60	1/30	1/15	1/8
f/stop:	2.8	4	5.6	8	11	16	22	32

The correlation of these shutter speeds and f /stops is 0.979. That sounds pretty high. You might assume that there must be a strong linear relationship. But when we check the scatterplot (we *always* check the scatterplot), it shows that something is not quite right:

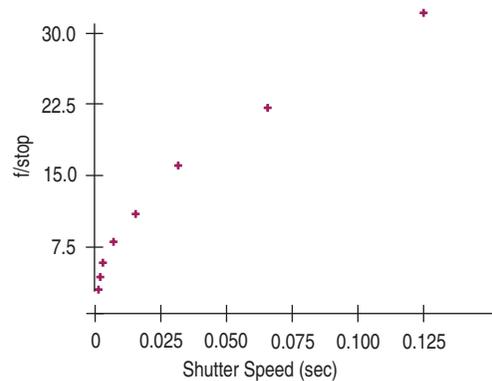


FIGURE 7.6

A scatterplot of f /stop vs. Shutter Speed shows a bent relationship.

We can see that the f /stop is not *linearly* related to the shutter speed. Can we find a transformation of f /stop that straightens out the line? What if we look at the *square* of the f /stop against the shutter speed?

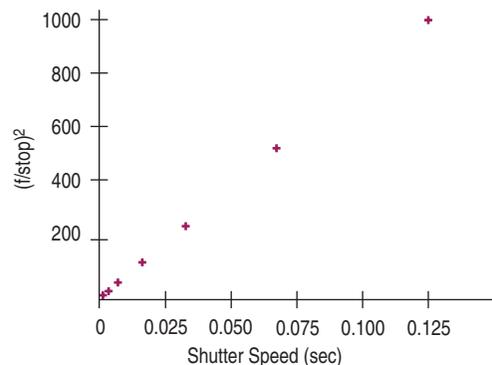


FIGURE 7.7

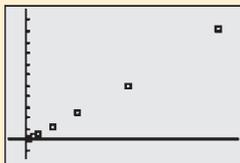
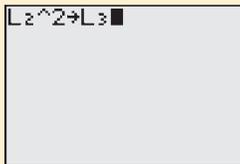
Re-expressing f /stop by squaring straightens the plot.

The second plot looks much more nearly straight. In fact, the correlation is now 0.998, but the increase in correlation is not important. (The original value of 0.979 should please almost anyone who sought a large correlation.) What is important is that the *form* of the plot is now straight, so the correlation is now an appropriate measure of association.¹⁰

We can often find transformations that straighten a scatterplot's form. Here, we found the square. Chapter 10 discusses simple ways to find a good re-expression.

¹⁰ Sometimes we can do a "reality check" on our choice of re-expression. In this case, a bit of research reveals that f /stops are related to the diameter of the open shutter. Since the amount of light that enters is determined by the *area* of the open shutter, which is related to the diameter by squaring, the square re-expression seems reasonable. Not all re-expressions have such nice explanations, but it's a good idea to think about them.

TI Tips



Straightening a curve

Let's straighten the f /stop scatterplot with your calculator.

- Enter the data in two lists, *shutterspeed* in **L1** and *f/stop* in **L2**.
- Set up a **STAT PLOT** to create a scatterplot with **Xlist:L1** and **Ylist:L2**.
- Hit **ZoomStat**. See the curve?

We want to find the squares of all the f /stops and save those re-expressed values in another datalist. That's easy to do.

- Create the command to square all the values in **L2** and **STO**re those results in **L3**, then hit **ENTER**.

Now make the new scatterplot.

- Go back to **STAT PLOT** and change the setup. **Xlist** is still **L1**, but this time specify **Ylist:L3**.
- **ZoomStat** again.

You now see the straightened plot for these data. On deck: drawing the best line through those points!

WHAT CAN GO WRONG?

Did you know that there's a strong correlation between playing an instrument and drinking coffee? No? One reason might be that the statement doesn't make sense. Correlation is a statistic that's valid only for *quantitative variables*.

- ▶ **Don't say "correlation" when you mean "association."** How often have you heard the word "correlation"? Chances are pretty good that when you've heard the term, it's been misused. When people want to sound scientific, they often say "correlation" when talking about the relationship between two variables. It's one of the most widely misused Statistics terms, and given how often statistics are misused, that's saying a lot. One of the problems is that many people use the specific term *correlation* when they really mean the more general term *association*. "Association" is a deliberately vague term describing the relationship between two variables.

"Correlation" is a precise term that measures the strength and direction of the linear relationship between quantitative variables.

- ▶ **Don't correlate categorical variables.** People who misuse the term "correlation" to mean "association" often fail to notice whether the variables they discuss are quantitative. Be sure to check the **Quantitative Variables Condition**.
- ▶ **Don't confuse correlation with causation.** One of the most common mistakes people make in interpreting statistics occurs when they observe a high correlation between two variables and jump to the perhaps tempting conclusion that one thing must be causing the other. Scatterplots and correlations *never* demonstrate causation. At best, these statistical tools can only reveal an association between variables, and that's a far cry from establishing cause and effect. While it's true that some associations may be causal, the nature and direction of the causation can be very hard to establish, and there's always the risk of overlooking lurking variables.
- ▶ **Make sure the association is linear.** Not all associations between quantitative variables are linear. Correlation can miss even a strong nonlinear association. A student project evaluating the quality of brownies baked at different temperatures reports a correlation of -0.05 between judges' scores and baking temperature. That seems to say there is no relationship—until we look at the scatterplot:

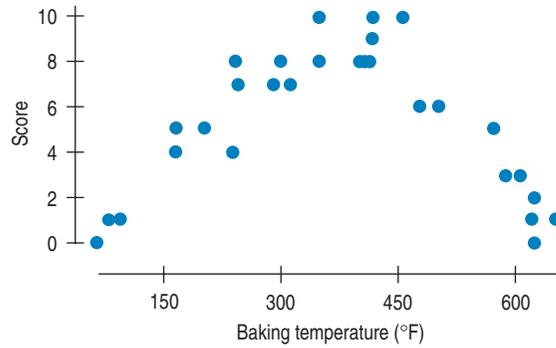


FIGURE 7.8
The relationship between brownie taste Score and Baking Temperature is strong, but not at all linear.

There is a strong association, but the relationship is not linear. Don't forget to check the Straight Enough Condition.

- ▶ **Don't assume the relationship is linear just because the correlation coefficient is high.** Recall that the correlation of f/stops and shutter speeds is 0.979 and yet the relationship is clearly not straight. Although the relationship must be straight for the correlation to be an appropriate measure, a high correlation is no guarantee of straightness. Nor is it safe to use correlation to judge the best re-expression. It's always important to look at the scatterplot.

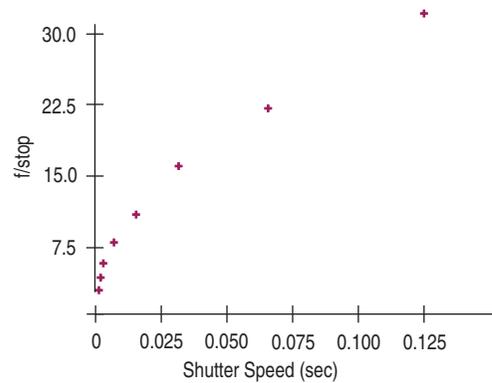
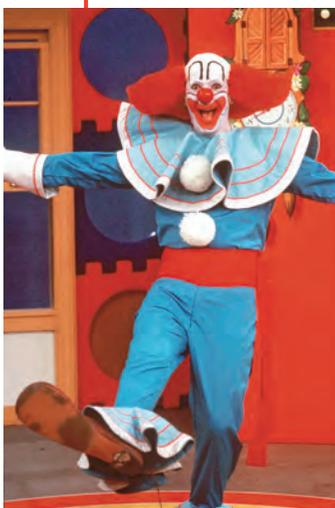


FIGURE 7.9
A scatterplot of f/stop vs. Shutter Speed shows a bent relationship even though the correlation is $r = 0.979$.



- ▶ **Beware of outliers.** You can't interpret a correlation coefficient safely without a background check for outliers. Here's a silly example:
The relationship between IQ and shoe size among comedians shows a surprisingly strong positive correlation of 0.50. To check assumptions, we look at the scatterplot:

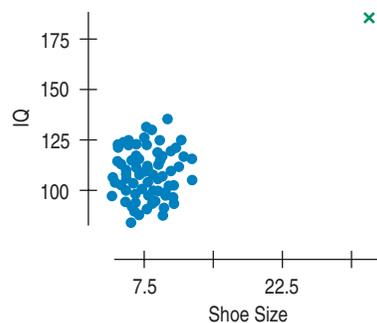
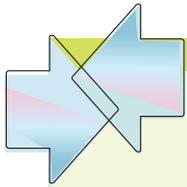


FIGURE 7.10
A scatterplot of IQ vs. Shoe Size. From this "study," what is the relationship between the two? The correlation is 0.50. Who does that point (the green x) in the upper right-hand corner belong to?

The outlier is Bozo the Clown, known for his large shoes, and widely acknowledged to be a comic "genius." Without Bozo, the correlation is near zero.

Even a single outlier can dominate the correlation value. That's why you need to check the Outlier Condition.





CONNECTIONS

Scatterplots are the basic tool for examining the relationship between two quantitative variables. We start with a picture when we want to understand the distribution of a single variable, and we always make a scatterplot to begin to understand the relationship between two quantitative variables.

We used z -scores as a way to measure the statistical distance of data values from their means. Now we've seen the z -scores of x and y working together to build the correlation coefficient. Correlation is a summary statistic like the mean and standard deviation—only it summarizes the strength of a linear relationship. And we interpret it as we did z -scores, using the standard deviations as our rulers in both x and y .

WHAT HAVE WE LEARNED?



A S **Simulation: Correlation, Center, and Scale.** If you have any lingering doubts that shifting and rescaling the data won't change the correlation, watch nothing happen right before your eyes!

In recent chapters we learned how to listen to the story told by data from a single variable. Now we've turned our attention to the more complicated (and more interesting) story we can discover in the association between two quantitative variables.

We've learned to begin our investigation by looking at a scatterplot. We're interested in the *direction* of the association, the *form* it takes, and its *strength*.

We've learned that, although not every relationship is linear, when the scatterplot is straight enough, the *correlation coefficient* is a useful numerical summary.

- ▶ The sign of the correlation tells us the direction of the association.
- ▶ The magnitude of the correlation tells us the *strength* of a linear association. Strong associations have correlations near -1 or $+1$ and very weak associations near 0 .
- ▶ Correlation has no units, so shifting or scaling the data, standardizing, or even swapping the variables has no effect on the numerical value.

Once again we've learned that doing Statistics right means we have to *Think* about whether our choice of methods is appropriate.

- ▶ The correlation coefficient is appropriate only if the underlying relationship is linear.
- ▶ We'll check the **Straight Enough Condition** by looking at a scatterplot.
- ▶ And, as always, we'll watch out for outliers!

Finally, we've learned not to make the mistake of assuming that a high correlation or strong association is evidence of a cause-and-effect relationship. Beware of lurking variables!

Terms

Scatterplots

147. A scatterplot shows the relationship between two quantitative variables measured on the same cases.

Association

- ▶ 147. **Direction:** A positive direction or association means that, in general, as one variable increases, so does the other. When increases in one variable generally correspond to decreases in the other, the association is negative.
- ▶ 147. **Form:** The form we care about most is straight, but you should certainly describe other patterns you see in scatterplots.
- ▶ 148. **Strength:** A scatterplot is said to show a strong association if there is little scatter around the underlying relationship.

Outlier

148. A point that does not fit the overall pattern seen in the scatterplot.

Response variable,
Explanatory variable,
x-variable, y-variable
Correlation Coefficient

149. In a scatterplot, you must choose a role for each variable. Assign to the y -axis the response variable that you hope to predict or explain. Assign to the x -axis the explanatory or predictor variable that accounts for, explains, predicts, or is otherwise responsible for the y -variable.

152. The correlation coefficient is a numerical measure of the direction and strength of a linear association.

$$r = \frac{\sum z_x z_y}{n - 1}$$

Lurking variable

157. A variable other than x and y that simultaneously affects both variables, accounting for the correlation between the two.

Skills

THINK

- ▶ Recognize when interest in the pattern of a possible relationship between two quantitative variables suggests making a scatterplot.
- ▶ Know how to identify the roles of the variables and that you should place the response variable on the y -axis and the explanatory variable on the x -axis.
- ▶ Know the conditions for correlation and how to check them.
- ▶ Know that correlations are between -1 and $+1$, and that each extreme indicates a perfect linear association.
- ▶ Understand how the magnitude of the correlation reflects the strength of a linear association as viewed in a scatterplot.
- ▶ Know that correlation has no units.
- ▶ Know that the correlation coefficient is not changed by changing the center or scale of either variable.

SHOW

- ▶ Understand that causation cannot be demonstrated by a scatterplot or correlation.
- ▶ Know how to make a scatterplot by hand (for a small set of data) or with technology.
- ▶ Know how to compute the correlation of two variables.

TELL

- ▶ Know how to read a correlation table produced by a statistics program.
- ▶ Be able to describe the direction, form, and strength of a scatterplot.
- ▶ Be prepared to identify and describe points that deviate from the overall pattern.
- ▶ Be able to use correlation as part of the description of a scatterplot.
- ▶ Be alert to misinterpretations of correlation.
- ▶ Understand that finding a correlation between two variables does not indicate a causal relationship between them. Beware the dangers of suggesting causal relationships when describing correlations.

SCATTERPLOTS AND CORRELATION ON THE COMPUTER

Statistics packages generally make it easy to look at a scatterplot to check whether the correlation is appropriate. Some packages make this easier than others.

Many packages allow you to modify or enhance a scatterplot, altering the axis labels, the axis numbering, the plot symbols, or the colors used. Some options, such as color and symbol choice, can be used to display additional information on the scatterplot.