CHAPTER 7

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

Hurricane Katrina killed 1,836 people[1] 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, NOAA[2] 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:

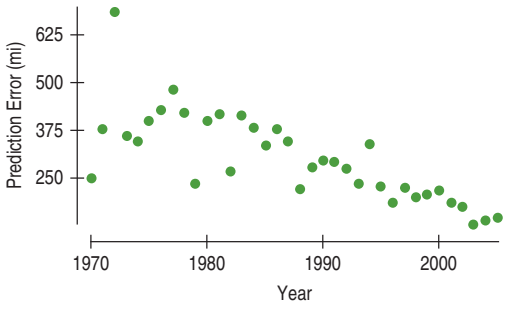


FIGURE 7.1
A scatterplot of the average error in nautical miles of the predicted position of Atlantic hurricanes for predictions made by the National Hurricane Center of NOAA, plotted against the Year *in which the predictions were made.*

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.

[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 Berra[3] 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

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.*

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

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.**

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

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.

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

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?

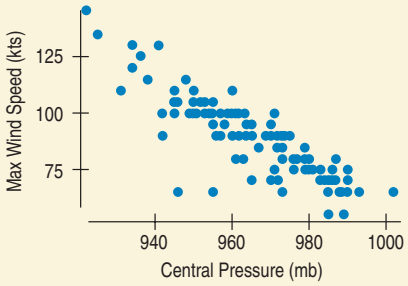
Finally, always look for the unexpected. Often the most interesting thing to see in a scatterplot is something you never thought to look for. One example of such a surprise is an **outlier** standing away from the overall pattern of the scatterplot. Such a point is almost always interesting and always 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.

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.

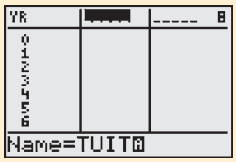
The roles that we choose for variables are more about how we *think* about them than about the variables themselves. Just placing a variable on the x-axis doesn't necessarily mean that it explains or predicts *anything*. And the variable on the y-axis may not respond to it in any way. We plotted prediction 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.

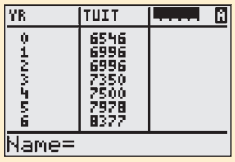
A S ***Self-Test:*** **Scatterplot Check.** Can you identify a scatterplot's direction, form, and strength?

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.

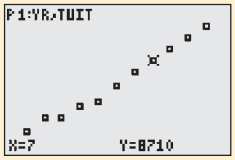




Naming the Lists

- 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.

[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.

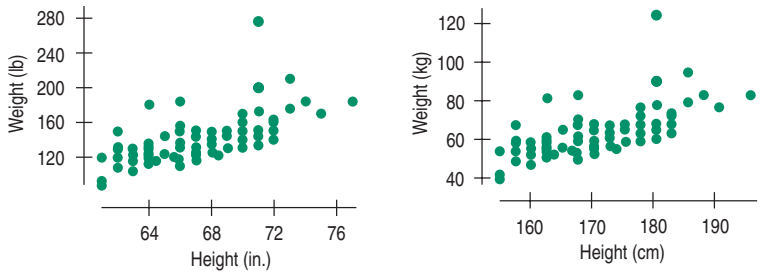


FIGURE 7.2 ***Weight vs. Height of Statistics students.***

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

A S ***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.

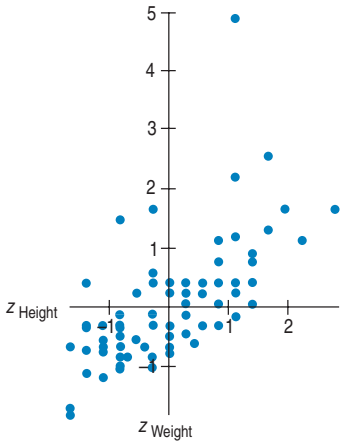
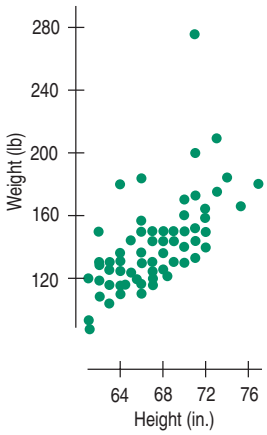
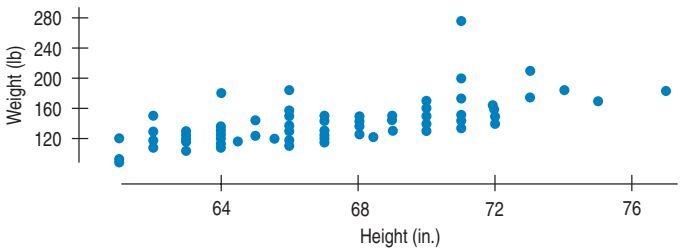


FIGURE 7.3
A scatterplot of standardized heights and weights.

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,

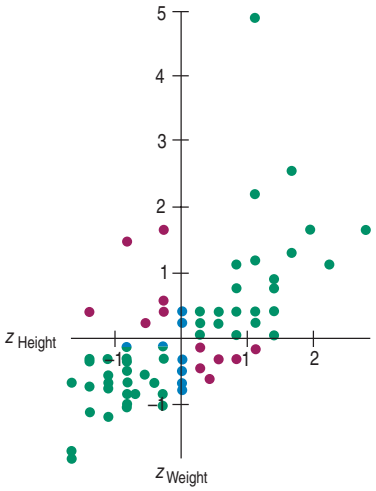


FIGURE 7.4

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.

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.

A S ***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.

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

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

[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.

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.

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

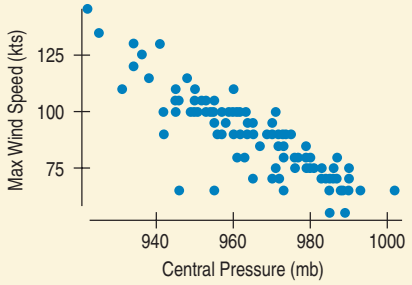
- **Outlier Condition:** Outliers can distort the correlation dramatically. An outlier can make an otherwise weak correlation look big or hide a strong correlation. It can even give an otherwise positive association a negative correlation coefficient (and vice versa). When you see an outlier, it's often 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

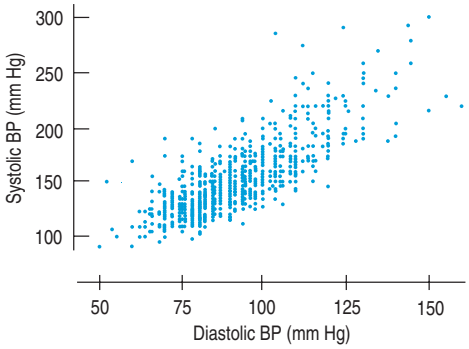
Plan State what you are trying to investigate.

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

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.

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]

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



Check the conditions.

- ✓ **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.

REALITY CHECK

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

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

SHOW

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

The correlation coefficient is $r = 0.792$.

[9] www.nhlbi.nih.gov/about/framingham

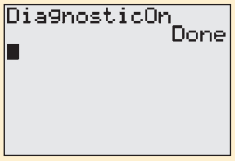
TELL

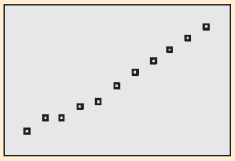
Conclusion Describe the direction, form, and strength you see in the plot, along with any unusual points or features. Be sure to state your interpretations in the proper context.

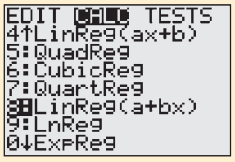
The scatterplot shows a positive direction, with higher SBP going with higher DBP. The plot is generally straight, with a moderate amount of scatter. The correlation of 0.792 is consistent with what I saw in the scatterplot. A few cases stand out with unusually high SBP compared with their DBP. It seems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

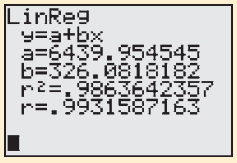
TI Tips

Finding the correlation









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) L4, L5

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

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.

A S ***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-*nspire*

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

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 ≠ 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.

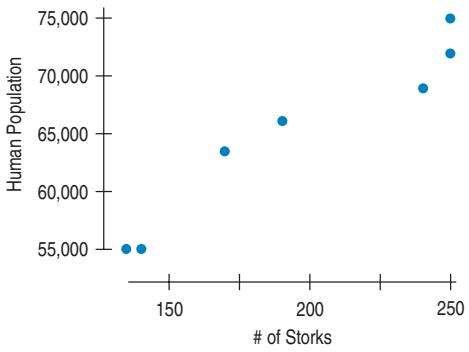


FIGURE 7.5
The number of storks in Oldenburg, Germany, plotted against the population of the town for 7 years in the 1930s. The association is clear. How about the causation? (Ornithologishe Monatsberichte, *44, no. 2)*

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	14	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:

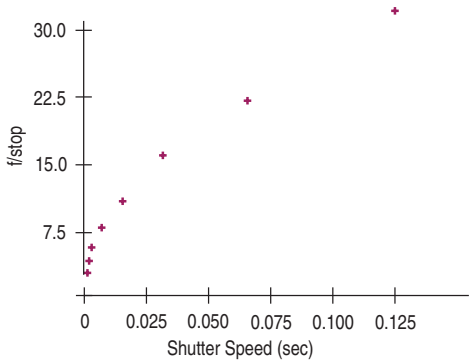


FIGURE 7.6
A scatterplot of f/stop *vs.* Shutter Speed *shows a bent relationship.*

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?

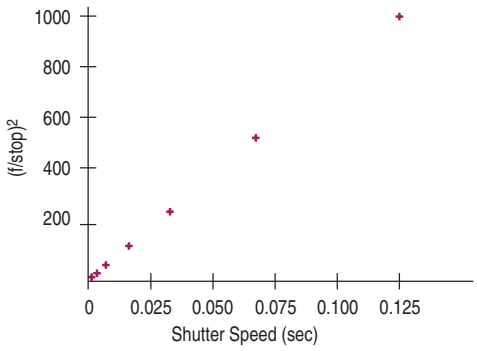


FIGURE 7.7
Re-expressing f/stop *by squaring straightens the plot.*

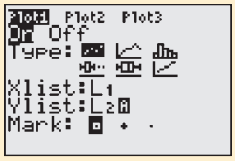
The second plot looks much more nearly straight. In fact, the correlation is now 0.998, but the increase in correlation is not important. (The original value of 0.979 should please almost anyone who sought a large correlation.) What is important is that the *form* of the plot is now straight, so the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

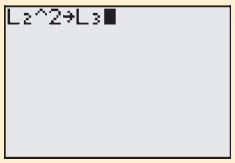
We can often find transformations that straighten a scatterplot's form. Here, we found the square. Chapter 10 discusses simple ways to find a good re-expression.

[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 STOre 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:

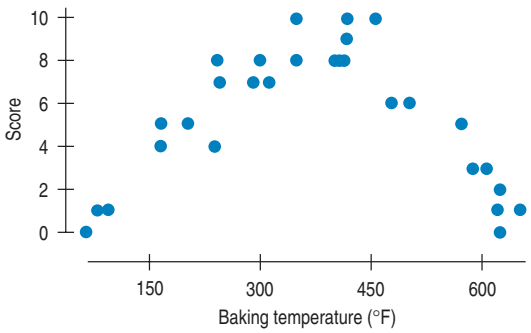


FIGURE 7.8
The relationship between brownie taste Score *and* Baking Temperature *is strong, but not at all linear.*

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.

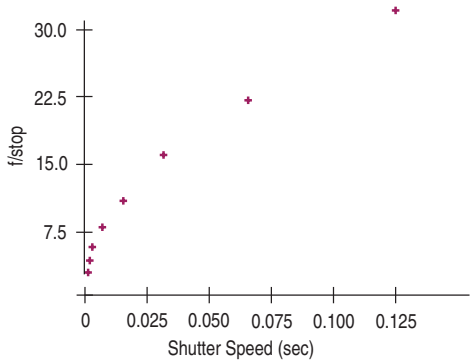


FIGURE 7.9
A scatterplot of f/stop *vs.* Shutter Speed *shows a bent relationship even though the correlation is* $r = 0.979$.

- **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:

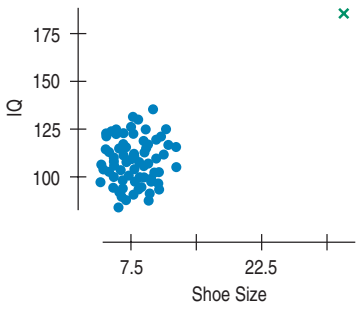


FIGURE 7.10
A scatterplot of IQ *vs.* Shoe Size. *From this "study," what is the relationship between the two? The correlation is 0.50. Who does that point (the green x) in the upper right-hand corner belong to?*

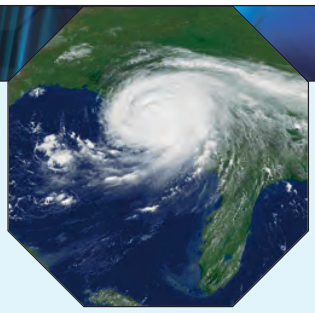
The outlier is Bozo the Clown, known for his large shoes, and widely acknowledged to be a comic "genius." Without Bozo, the correlation is near zero.

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?

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!

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!

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

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.

Correlation Coefficient

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

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

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.
- Understand that causation cannot be demonstrated by a scatterplot or correlation.

SHOW

- 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.
- Know how to read a correlation table produced by a statistics program.

TELL

- 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.

EXERCISES

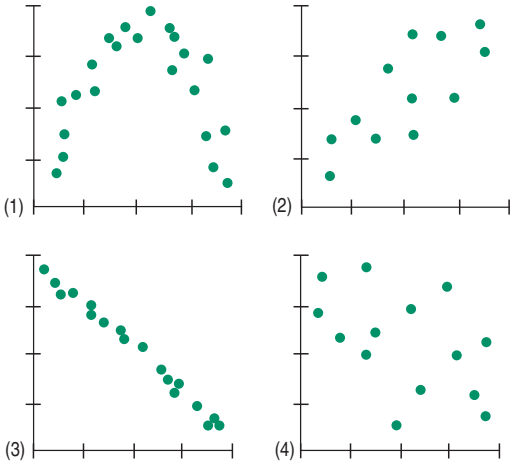
1. **Association.** Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.
   a) Apples: weight in grams, weight in ounces
   b) Apples: circumference (inches), weight (ounces)
   c) College freshmen: shoe size, grade point average
   d) Gasoline: number of miles you drove since filling up, gallons remaining in your tank

2. **Association.** Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.
   a) T-shirts at a store: price each, number sold
   b) Scuba diving: depth, water pressure
   c) Scuba diving: depth, visibility
   d) All elementary school students: weight, score on a reading test

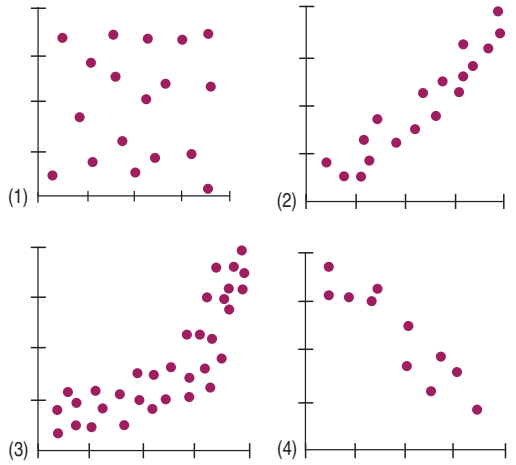
3. **Association.** Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.
   a) When climbing mountains: altitude, temperature
   b) For each week: ice cream cone sales, air-conditioner sales
   c) People: age, grip strength
   d) Drivers: blood alcohol level, reaction time

4. **Association.** Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.
   a) Long-distance calls: time (minutes), cost
   b) Lightning strikes: distance from lightning, time delay of the thunder
   c) A streetlight: its apparent brightness, your distance from it
   d) Cars: weight of car, age of owner

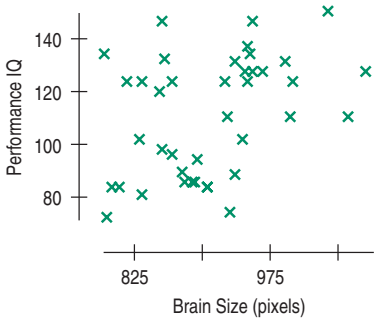
5. **Scatterplots.** Which of the scatterplots at the top of the next column show
   a) little or no association?
   b) a negative association?
   c) a linear association?
   d) a moderately strong association?
   e) a very strong association?



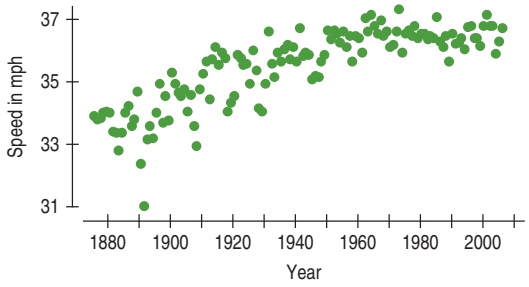
6. **Scatterplots.** Which of the scatterplots below show
   a) little or no association?
   b) a negative association?
   c) a linear association?
   d) a moderately strong association?
   e) a very strong association?



T 7. **Performance IQ scores vs. brain size.** A study examined brain size (measured as pixels counted in a digitized magnetic resonance image [MRI] of a cross section of the brain) and IQ (4 Performance scales of the Weschler IQ test) for college students. The scatterplot shows the Performance IQ scores vs. the brain size. Comment on the association between brain size and IQ as seen in the scatterplot on the next page.

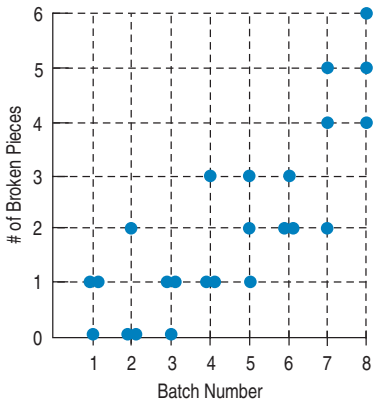


T **8. Kentucky Derby 2006.** The fastest horse in Kentucky Derby history was Secretariat in 1973. The scatterplot shows speed (in miles per hour) of the winning horses each year.



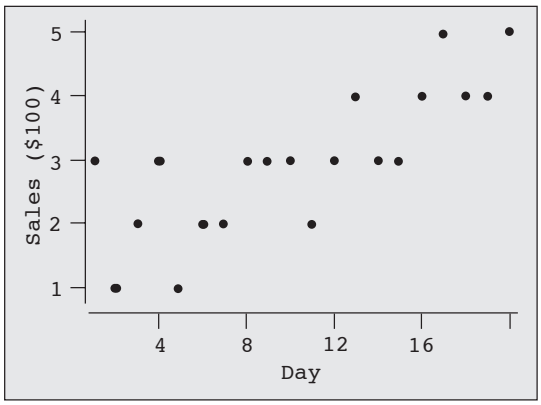
What do you see? In most sporting events, performances have improved and continue to improve, so surely we anticipate a positive direction. But what of the form? Has the performance increased at the same rate throughout the last 130 years?

9. Firing pottery. A ceramics factory can fire eight large batches of pottery a day. Sometimes a few of the pieces break in the process. In order to understand the problem better, the factory records the number of broken pieces in each batch for 3 days and then creates the scatterplot shown.



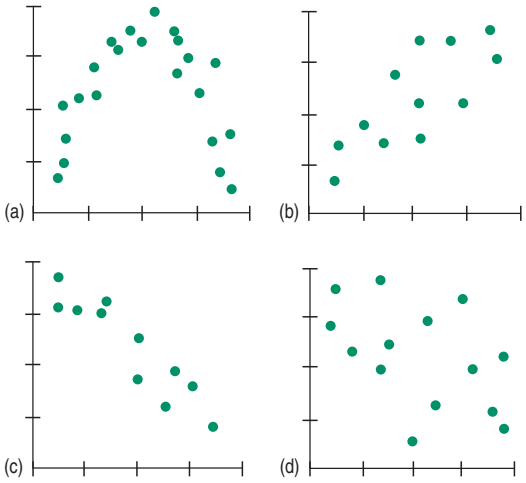
a) Make a histogram showing the distribution of the number of broken pieces in the 24 batches of pottery examined.
b) Describe the distribution as shown in the histogram. What feature of the problem is more apparent in the histogram than in the scatterplot?
c) What aspect of the company's problem is more apparent in the scatterplot?

10. Coffee sales. Owners of a new coffee shop tracked sales for the first 20 days and displayed the data in a scatterplot (by day).

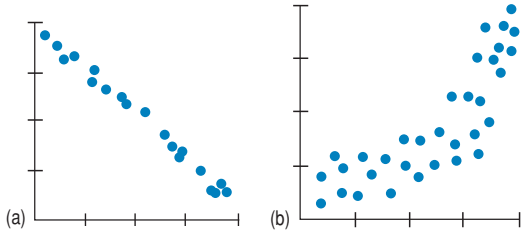


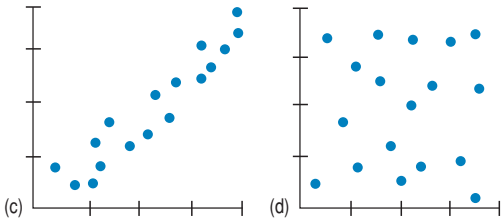
a) Make a histogram of the daily sales since the shop has been in business.
b) State one fact that is obvious from the scatterplot, but not from the histogram.
c) State one fact that is obvious from the histogram, but not from the scatterplot.

11. Matching. Here are several scatterplots. The calculated correlations are −0.923, −0.487, 0.006, and 0.777. Which is which?



12. Matching. Here and on the next page are several scatterplots. The calculated correlations are −0.977, −0.021, 0.736, and 0.951. Which is which?

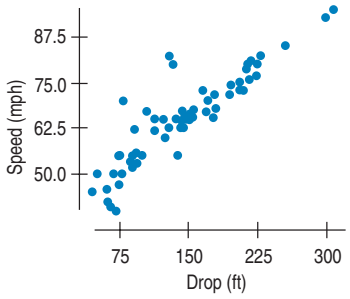




13. Politics. A candidate for office claims that "there is a correlation between television watching and crime." Criticize this statement on statistical grounds.

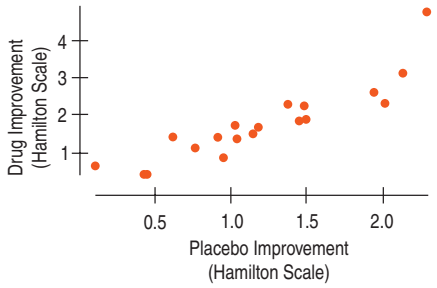
14. Car thefts. The National Insurance Crime Bureau reports that Honda Accords, Honda Civics, and Toyota Camrys are the cars most frequently reported stolen, while Ford Tauruses, Pontiac Vibes, and Buick LeSabres are stolen least often. Is it reasonable to say that there's a correlation between the type of car you own and the risk that it will be stolen?

T **15. Roller coasters.** Roller coasters get all their speed by dropping down a steep initial incline, so it makes sense that the height of that drop might be related to the speed of the coaster. Here's a scatterplot of top *Speed* and largest *Drop* for 75 roller coasters around the world.



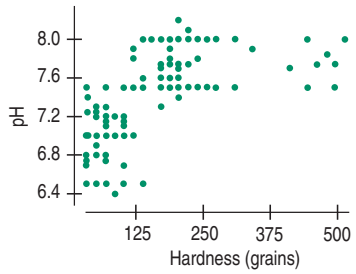
a) Does the scatterplot indicate that it is appropriate to calculate the correlation? Explain.
b) In fact, the correlation of *Speed* and *Drop* is 0.91. Describe the association.

T **16. Antidepressants.** A study compared the effectiveness of several antidepressants by examining the experiments in which they had passed the FDA requirements. Each of those experiments compared the active drug with a placebo, an inert pill given to some of the subjects. In each experiment some patients treated with the placebo had improved, a phenomenon called the *placebo effect*. Patients' depression levels were evaluated on the Hamilton Depression Rating Scale, where larger numbers indicate greater improvement. (The Hamilton scale is a widely accepted standard that was used in each of the independently run studies.) The scatterplot at the top of the next column compares mean improvement levels for the antidepressants and placebos for several experiments.



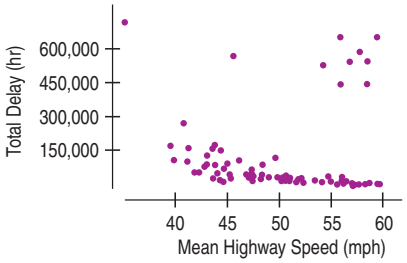
a) Is it appropriate to calculate the correlation? Explain.
b) The correlation is 0.898. Explain what we have learned about the results of these experiments.

T **17. Hard water.** In a study of streams in the Adirondack Mountains, the following relationship was found between the water's pH and its hardness (measured in grains):



Is it appropriate to summarize the strength of association with a correlation? Explain.

18. Traffic headaches. A study of traffic delays in 68 U.S. cities found the following relationship between total delays (in total hours lost) and mean highway speed:



Is it appropriate to summarize the strength of association with a correlation? Explain.

19. Cold nights. Is there an association between time of year and the nighttime temperature in North Dakota? A researcher assigned the numbers 1–365 to the days January 1–December 31 and recorded the temperature at 2:00 a.m. for each. What might you expect the correlation between *DayNumber* and *Temperature* to be? Explain.

20. Association. A researcher investigating the association between two variables collected some data and was surprised when he calculated the correlation. He had expected to find a fairly strong association, yet the correlation was near 0. Discouraged, he didn't bother making a scatterplot. Explain to him how the scatterplot could still reveal the strong association he anticipated.

21. Prediction units. The errors in predicting hurricane tracks (examined in this chapter) were given in nautical miles. An ordinary mile is 0.86898 nautical miles. Most people living on the Gulf Coast of the United States would prefer to know the prediction errors in miles rather than nautical miles. Explain why converting the errors to miles would not change the correlation between *Prediction Error* and *Year.*

22. More predictions. Hurricane Katrina's hurricane force winds extended 120 miles from its center. Katrina was a big storm, and that affects how we think about the prediction errors. Suppose we add 120 miles to each error to get an idea of how far from the predicted track we might still find damaging winds. Explain what would happen to the correlation between *Prediction Error* and *Year,* and why.

23. Correlation errors. Your Economics instructor assigns your class to investigate factors associated with the gross domestic product (*GDP*) of nations. Each student examines a different factor (such as *Life Expectancy, Literacy Rate,* etc.) for a few countries and reports to the class. Apparently, some of your classmates do not understand Statistics very well because you know several of their conclusions are incorrect. Explain the mistakes in their statements below.

a) "My very low correlation of −0.772 shows that there is almost no association between *GDP* and *Infant Mortality Rate.*"

b) "There was a correlation of 0.44 between *GDP* and *Continent.*"

24. More correlation errors. Students in the Economics class discussed in Exercise 23 also wrote these conclusions. Explain the mistakes they made.

a) "There was a very strong correlation of 1.22 between *Life Expectancy* and *GDP.*"

b) "The correlation between *Literacy Rate* and *GDP* was 0.83. This shows that countries wanting to increase their standard of living should invest heavily in education."

25. Height and reading. A researcher studies children in elementary school and finds a strong positive linear association between height and reading scores.

a) Does this mean that taller children are generally better readers?

b) What might explain the strong correlation?

26. Cellular telephones and life expectancy. A survey of the world's nations in 2004 shows a strong positive correlation between percentage of the country using cell phones and life expectancy in years at birth.

a) Does this mean that cell phones are good for your health?

b) What might explain the strong correlation?

27. Correlation conclusions I. The correlation between *Age* and *Income* as measured on 100 people is $r = 0.75$. Explain whether or not each of these possible conclusions is justified:

a) When *Age* increases, *Income* increases as well.

b) The form of the relationship between *Age* and *Income* is straight.

c) There are no outliers in the scatterplot of *Income* vs. *Age.*

d) Whether we measure *Age* in years or months, the correlation will still be 0.75.

28. Correlation conclusions II. The correlation between *Fuel Efficiency* (as measured by miles per gallon) and *Price* of 150 cars at a large dealership is $r = -0.34$. Explain whether or not each of these possible conclusions is justified:

a) The more you pay, the lower the fuel efficiency of your car will be.

b) The form of the relationship between *Fuel Efficiency* and *Price* is moderately straight.

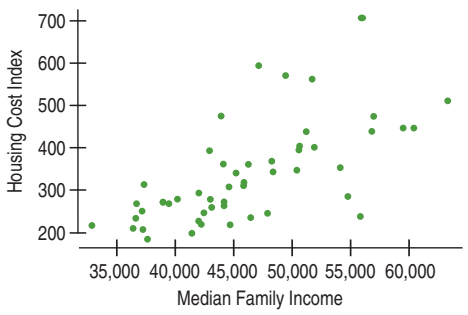
c) There are several outliers that explain the low correlation.

d) If we measure *Fuel Efficiency* in kilometers per liter instead of miles per gallon, the correlation will increase.

29. Baldness and heart disease. Medical researchers followed 1435 middle-aged men for a period of 5 years, measuring the amount of *Baldness* present (none = 1, little = 2, some = 3, much = 4, extreme = 5) and presence of *Heart Disease* (No = 0, Yes = 1). They found a correlation of 0.089 between the two variables. Comment on their conclusion that this shows that baldness is not a possible cause of heart disease.

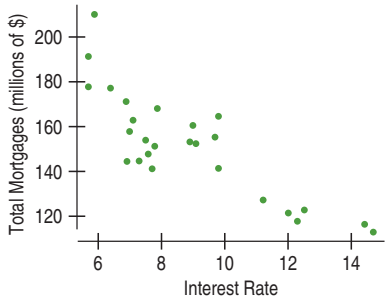
30. Sample survey. A polling organization is checking its database to see if the two data sources it used sampled the same zip codes. The variable *Datasource* = 1 if the data source is MetroMedia, 2 if the data source is DataQwest, and 3 if it's RollingPoll. The organization finds that the correlation between five-digit zip code and *Datasource* is −0.0229. It concludes that the correlation is low enough to state that there is no dependency between *Zip Code* and *Source of Data.* Comment.

T **31. Income and housing.** The Office of Federal Housing Enterprise Oversight (www.ofheo.gov) collects data on various aspects of housing costs around the United States. Here is a scatterplot of the *Housing Cost Index* versus the *Median Family Income* for each of the 50 states. The correlation is 0.65.



a) Describe the relationship between the *Housing Cost Index* and the *Median Family Income* by state.
b) If we standardized both variables, what would the correlation coefficient between the standardized variables be?
c) If we had measured *Median Family Income* in thousands of dollars instead of dollars, how would the correlation change?
d) Washington, DC, has a Housing Cost Index of 548 and a median income of about $45,000. If we were to include DC in the data set, how would that affect the correlation coefficient?
e) Do these data provide proof that by raising the median income in a state, the Housing Cost Index will rise as a result? Explain.

T **32. Interest rates and mortgages.** Since 1980, average mortgage interest rates have fluctuated from a low of under 6% to a high of over 14%. Is there a relationship between the amount of money people borrow and the interest rate that's offered? Here is a scatterplot of *Total Mortgages* in the United States (in millions of 2005 dollars) versus *Interest Rate* at various times over the past 26 years. The correlation is −0.84.



a) Describe the relationship between *Total Mortgages* and *Interest Rate*.
b) If we standardized both variables, what would the correlation coefficient between the standardized variables be?
c) If we were to measure *Total Mortgages* in thousands of dollars instead of millions of dollars, how would the correlation coefficient change?
d) Suppose in another year, interest rates were 11% and mortgages totaled $250 million. How would including that year with these data affect the correlation coefficient?
e) Do these data provide proof that if mortgage rates are lowered, people will take out more mortgages? Explain.

T **33. Fuel economy 2007.** Here are advertised horsepower ratings and expected gas mileage for several 2007 vehicles. (http://www.kbb.com/KBB/ReviewsAndRatings)

Vehicle	Horsepower	Highway Gas Mileage (mpg)
Audi A4	200	32
BMW 328	230	30
Buick LaCrosse	200	30
Chevy Cobalt	148	32
Chevy TrailBlazer	291	22
Ford Expedition	300	20
GMC Yukon	295	21
Honda Civic	140	40
Honda Accord	166	34
Hyundai Elantra	138	36
Lexus IS 350	306	28
Lincoln Navigator	300	18
Mazda Tribute	212	25
Toyota Camry	158	34
Volkswagen Beetle	150	30

a) Make a scatterplot for these data.
b) Describe the direction, form, and strength of the plot.
c) Find the correlation between horsepower and miles per gallon.
d) Write a few sentences telling what the plot says about fuel economy.

34. Drug abuse. A survey was conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. The results are summarized in the table.

	Percent Who Have Used	
Country	**Marijuana**	**Other Drugs**
Czech Rep.	22	4
Denmark	17	3
England	40	21
Finland	5	1
Ireland	37	16
Italy	19	8
No. Ireland	23	14
Norway	6	3
Portugal	7	3
Scotland	53	31
USA	34	24

a) Create a scatterplot.
b) What is the correlation between the percent of teens who have used marijuana and the percent who have used other drugs?
c) Write a brief description of the association.
d) Do these results confirm that marijuana is a "gateway drug," that is, that marijuana use leads to the use of other drugs? Explain.

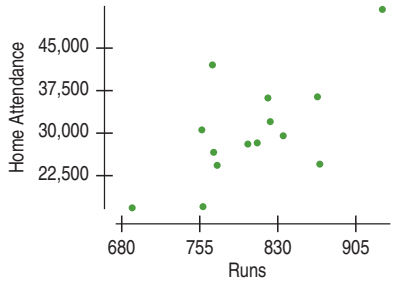
35. Burgers. Fast food is often considered unhealthy because much of it is high in both fat and sodium. But are the two related? Here are the fat and sodium contents of several brands of burgers. Analyze the association between fat content and sodium.

Fat (g)	19	31	34	35	39	39	43
Sodium (mg)	920	1500	1310	860	1180	940	1260

36. Burgers II. In the previous exercise you analyzed the association between the amounts of fat and sodium in fast food hamburgers. What about fat and calories? Here are data for the same burgers:

Fat (g)	19	31	34	35	39	39	43
Calories	410	580	590	570	640	680	660

37. Attendance 2006. American League baseball games are played under the designated hitter rule, meaning that pitchers, often weak hitters, do not come to bat. Baseball owners believe that the designated hitter rule means more runs scored, which in turn means higher attendance. Is there evidence that more fans attend games if the teams score more runs? Data collected from American League games during the 2006 season indicate a correlation of 0.667 between runs scored and the number of people at the game. (http://mlb.mlb.com)

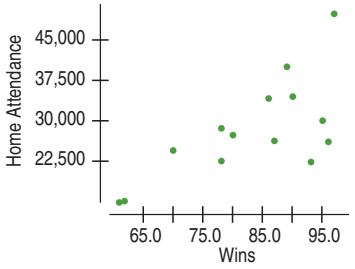


a) Does the scatterplot indicate that it's appropriate to calculate a correlation? Explain.
b) Describe the association between attendance and runs scored.
c) Does this association prove that the owners are right that more fans will come to games if the teams score more runs?

38. Second inning 2006. Perhaps fans are just more interested in teams that win. The displays below are based on American League teams for the 2006 season. (http://espn.go.com) Are the teams that win necessarily those which score the most runs?

CORRELATION

	Wins	Runs	Attend
Wins	1.000		
Runs	0.605	1.000	
Attend	0.697	0.667	1.000



a) Do winning teams generally enjoy greater attendance at their home games? Describe the association.
b) Is attendance more strongly associated with winning or scoring runs? Explain.
c) How strongly is scoring more runs associated with winning more games?

39. Thrills. People who responded to a July 2004 Discovery Channel poll named the 10 best roller coasters in the United States. The table below shows the length of the initial drop (in feet) and the duration of the ride (in seconds). What do these data indicate about the height of a roller coaster and the length of the ride you can expect?

Roller Coaster	State	Drop (ft)	Duration (sec)
Incredible Hulk	FL	105	135
Millennium Force	OH	300	105
Goliath	CA	255	180
Nitro	NJ	215	240
Magnum XL-2000	OH	195	120
The Beast	OH	141	65
Son of Beast	OH	214	140
Thunderbolt	PA	95	90
Ghost Rider	CA	108	160
Raven	IN	86	90

40. Vehicle weights. The Minnesota Department of Transportation hoped that they could measure the weights of big trucks without actually stopping the vehicles by using a newly developed "weight-in-motion" scale. To see if the new device was accurate, they conducted a calibration test. They weighed several stopped trucks (static weight) and assumed that this weight was correct. Then they weighed the trucks again while they were moving to see how well the new scale could estimate the actual weight. Their data are given in the table on the next page.

WEIGHTS (1000S OF LBS)

Weight-in-Motion	Static Weight
26.0	27.9
29.9	29.1
39.5	38.0
25.1	27.0
31.6	30.3
36.2	34.5
25.1	27.8
31.0	29.6
35.6	33.1
40.2	35.5

a) Make a scatterplot for these data.
b) Describe the direction, form, and strength of the plot.
c) Write a few sentences telling what the plot says about the data. (*Note*: The sentences should be about weighing trucks, not about scatterplots.)
d) Find the correlation.
e) If the trucks were weighed in kilograms, how would this change the correlation? (1 kilogram = 2.2 pounds)
f) Do any points deviate from the overall pattern? What does the plot say about a possible recalibration of the weight-in-motion scale?

41. Planets (more or less). On August 24, 2006, the International Astronomical Union voted that Pluto is not a planet. Some members of the public have been reluctant to accept that decision. Let's look at some of the data. (We'll see more in the next chapter.) Is there any pattern to the locations of the planets? The table shows the average distance of each of the traditional nine planets from the sun.

Planet	Position Number	Distance from Sun (million miles)
Mercury	1	36
Venus	2	67
Earth	3	93
Mars	4	142
Jupiter	5	484
Saturn	6	887
Uranus	7	1784
Neptune	8	2796
Pluto	9	3666

a) Make a scatterplot and describe the association. (Remember: direction, form, and strength!)
b) Why would you not want to talk about the correlation between a planet's *Position* and *Distance* from the sun?
c) Make a scatterplot showing the logarithm of *Distance* vs. *Position*. What is better about this scatterplot?

42. Flights. The number of flights by U.S. Airlines has grown rapidly. Here are the number of flights flown in each year from 1995 to 2005.
a) Find the correlation of *Flights* with *Year*.
b) Make a scatterplot and describe the trend.
c) Note two reasons that the correlation you found in (a) is not a suitable summary of the strength of the association. Can you account for these violations of the conditions?

Year	Flights
1995	5,327,435
1996	5,351,983
1997	5,411,843
1998	5,384,721
1999	5,527,884
2000	5,683,047
2001	5,967,780
2002	5,271,359
2003	6,488,539
2004	7,129,270
2005	7,140,596

JUST CHECKING Answers

1. We know the scores are quantitative. We should check to see if the Straight Enough Condition and the Outlier Condition are satisfied by looking at a scatterplot of the two scores.
2. It won't change.
3. It won't change.
4. They are likely to have done poorly. The positive correlation means that low scores on Exam 1 are associated with low scores on Exam 2 (and similarly for high scores).
5. No. The general association is positive, but individual performances may vary.