Chapter 14

Describing Relationships:

Scatterplots and Correlation

Lecture Slides

Case Study: Describing Relationships – Scatterplots and Correlation 1

The news media have a weakness for lists.

- Best places to live
- Best colleges
- Healthiest foods
- Worst-dressed women

A list of best or worst is sure to find a place in the news.

Case Study: Describing Relationships – Scatterplots and Correlation 2

When the state-by-state SAT scores come out each year, it's therefore no surprise that we find news articles ranking the states from best (Minnesota) to worst (District of Columbia) according to the average SAT Mathematics score achieved by their high school seniors.

Such reports leave readers believing that schools in the District of Columbia must be much worse than those in Minnesota.

Case Study: Describing Relationships – Scatterplots and Correlation 3

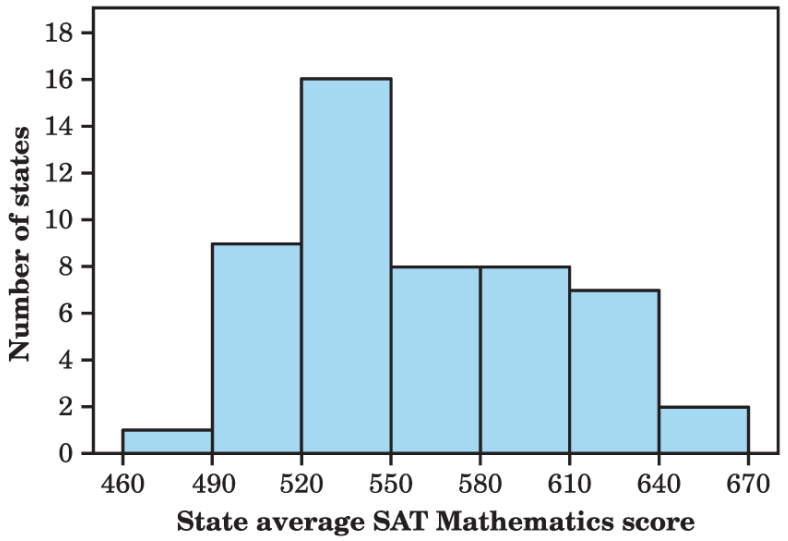
The College Board, which sponsors the SAT exams, doesn't like this practice at all.

"Comparing or ranking states on the basis of SAT scores alone is invalid and strongly discouraged by the College Board," says the heading on their table of state average SAT scores.

To see why, let's look at the data.

Case Study: Describing Relationships – Scatterplots and Correlation 4

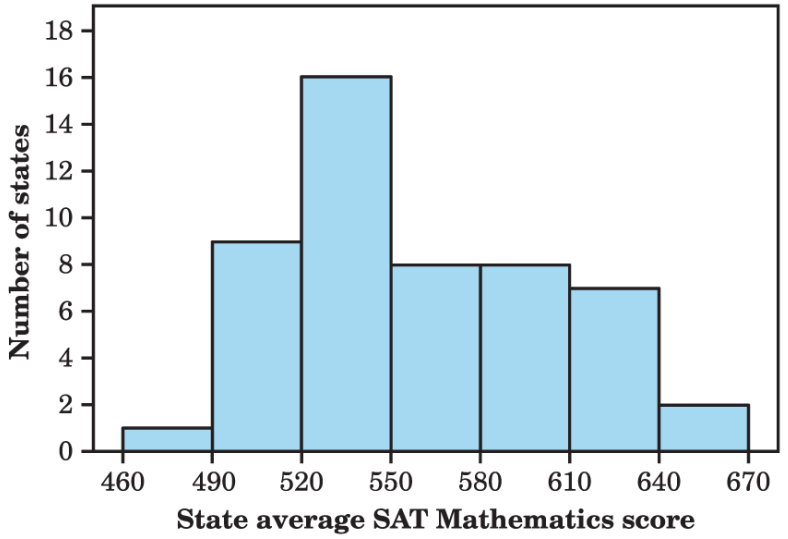
Figure 14.1 shows the distribution of average scores on the SAT Mathematics exam for the 50 states and the District of Columbia. Minnesota leads at 651, and the District of Columbia trails at 468 on the SAT scale of 200 to 800.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Case Study: Describing Relationships – Scatterplots and Correlation 5

The distribution has one clear peak and is skewed to the right.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Case Study: Describing Relationships – Scatterplots and Correlation 6

In this chapter we will learn that to understand one variable, such as SAT scores, we must look at how it is related to other variables.

By the end of this chapter you will be able to use what you have learned to appreciate why the College Board discourages ranking states on SAT scores alone.

Relationships among Variables 1

A medical study finds that short women are more likely to have heart attacks than women of average height, while tall women have the fewest heart attacks.

An insurance group reports that heavier cars are involved in fewer fatal accidents per 10,000 vehicles registered than are lighter cars.

These and many other statistical studies look at the relationship between two variables.

The relationship between two variables can be strongly influenced by other variables that are lurking in the background.

Relationships among Variables 2

To examine the relationship between two variables:

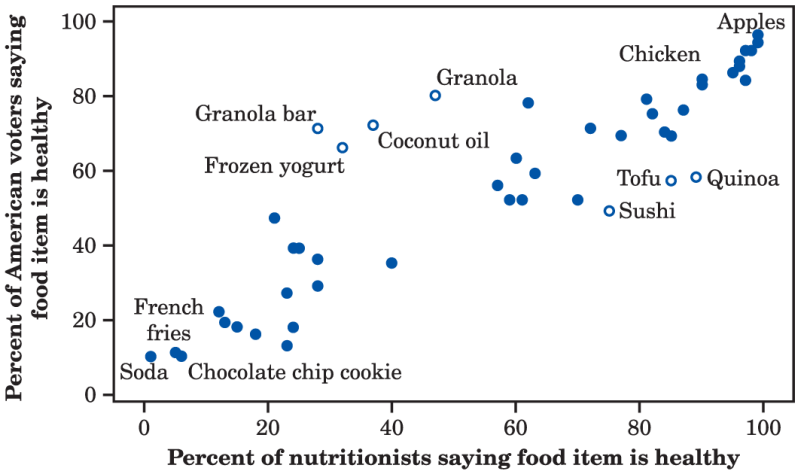
- First plot the data, then add numerical summaries.
- Look for overall patterns and deviations from those patterns.
- When the overall pattern is quite regular, there is sometimes a way to describe it very briefly.

Example: Food Nutrition

Figure 14.2 is a scatterplot that shows percent of nutritionists saying a food item is healthy is related to the percent of American voters saying a food item is healthy.

"Percent of nutritionists saying food item is healthy" is the explanatory variable.

"Percent of American voters saying food item is healthy" is the response variable.

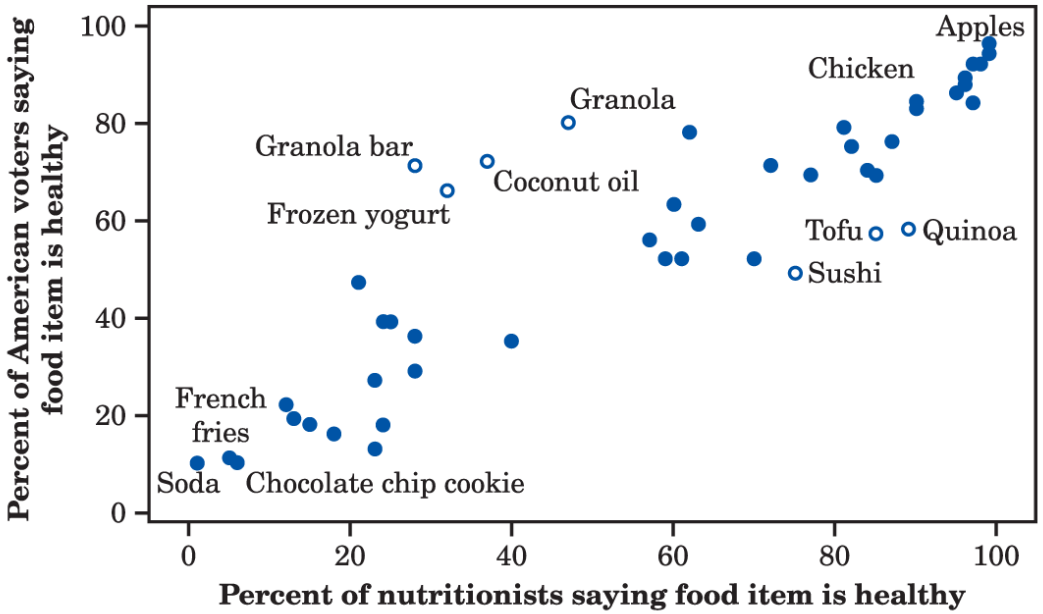


Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Food Nutrition 2

Each point represents one food item.

Generally, as the percent of nutritionists saying the food item is healthy increases, the percent of American voters saying the food item is healthy also increases.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Scatterplots

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Always plot the explanatory variable, if there is one, on the horizontal axis (the *x*-axis) of a scatterplot.

If there is no explanatory-response distinction, either variable can go on the horizontal axis.

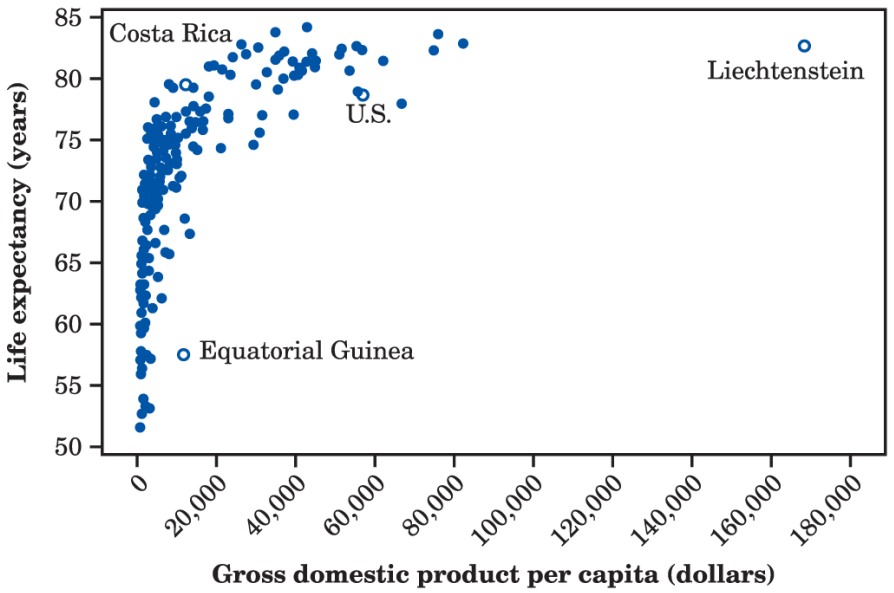
Example: Health and wealth 1

Figure 14.3 is a scatterplot of data from the World Bank for 2016.

The individuals are all the world's nations for which data are available.

The explanatory variable is the gross domestic product (GDP) per capita.

The response variable is life expectancy at birth.

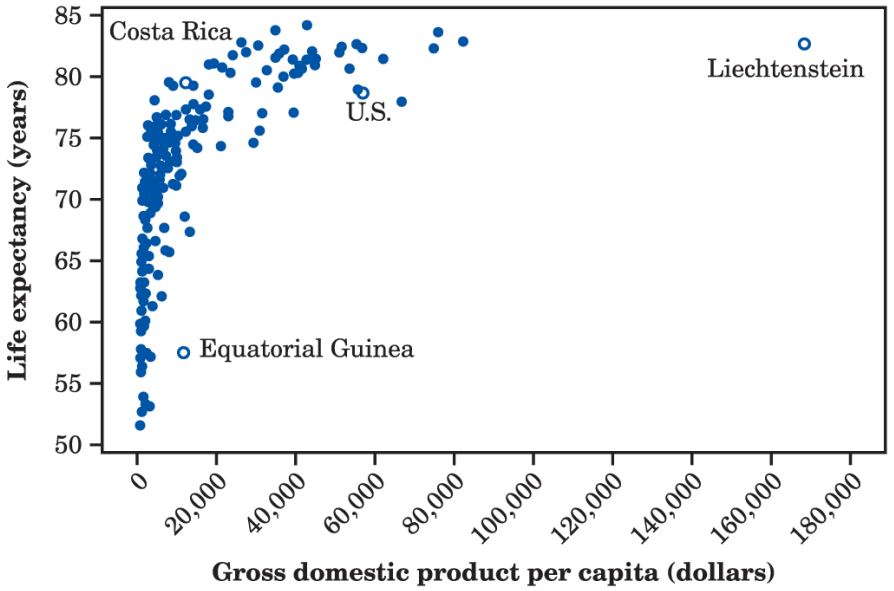


Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Health and wealth 2

Life expectancy tends to rise very quickly as GDP increases and then levels off.

People in very rich countries such as the United States typically live no longer than people in poorer, but not extremely poor, nations. Some of these countries, such as Costa Rica, do almost as well as the United States.



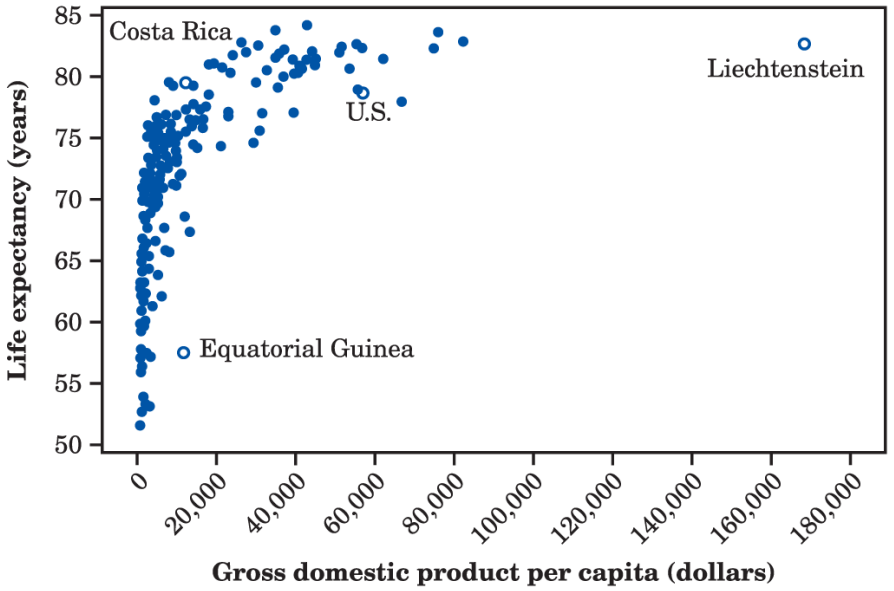
Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Health and wealth 3

Two nations are outliers.

In Equatorial Guinea, life expectancies are similar to those of its neighbors, but its GDP is higher.

Equatorial Guinea produces oil. It may be that income from mineral exports goes mainly to a few people, pulling up GDP per capita without affecting income or life expectancy of ordinary citizens.

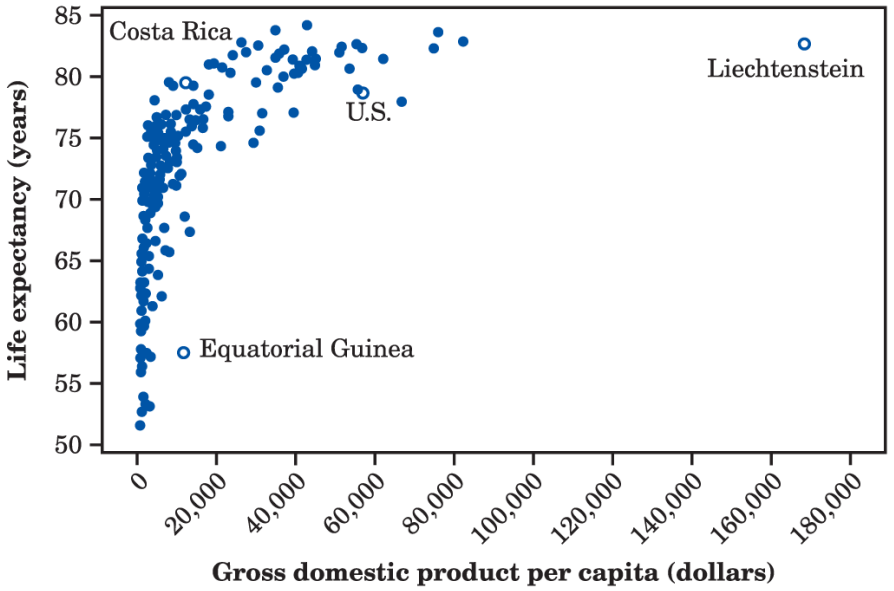


Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Health and wealth 4

The other outlier is Liechtenstein, a tiny nation bordering Switzerland and Austria.

Liechtenstein has a strong financial sector and is considered a tax haven.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Interpreting Scatterplots 1

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a scatterplot by the **direction, form, and strength** of the relationship.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Interpreting Scatterplots 2

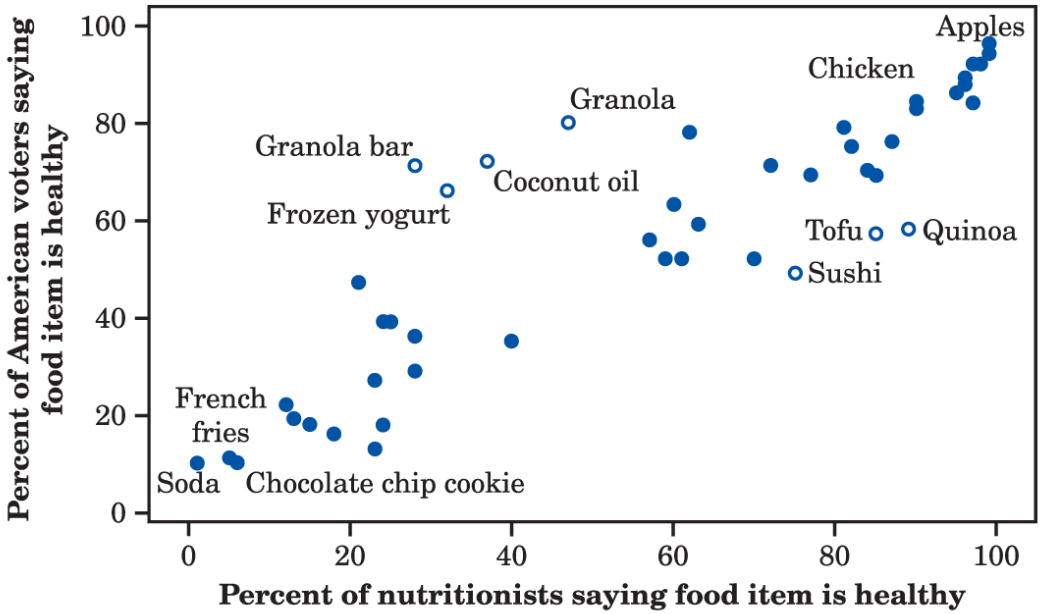
Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together. The scatterplot slopes upward as we move from left to right.

Two variables are **negatively associated** when above-average values of one tend to accompany below-average values of the other, and vice versa. The scatterplot slopes downward from left to right.

Example: Scatterplots 1

The form on Figure 14.2 is roughly a straight line. This is not a strong relationship.

There is a positive association.

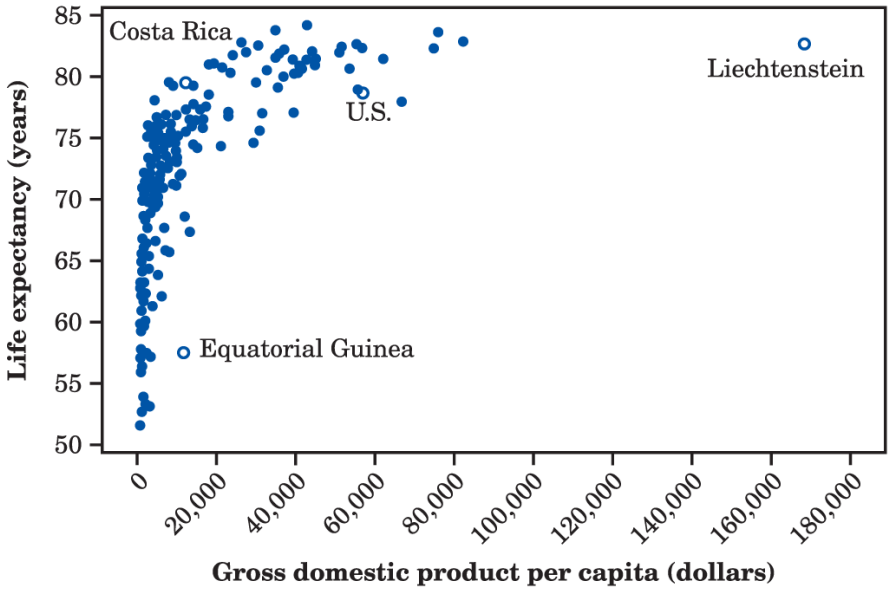


Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Scatterplots 2

The form on Figure 14.3 is curved. This is not a strong relationship.

There is a positive association.

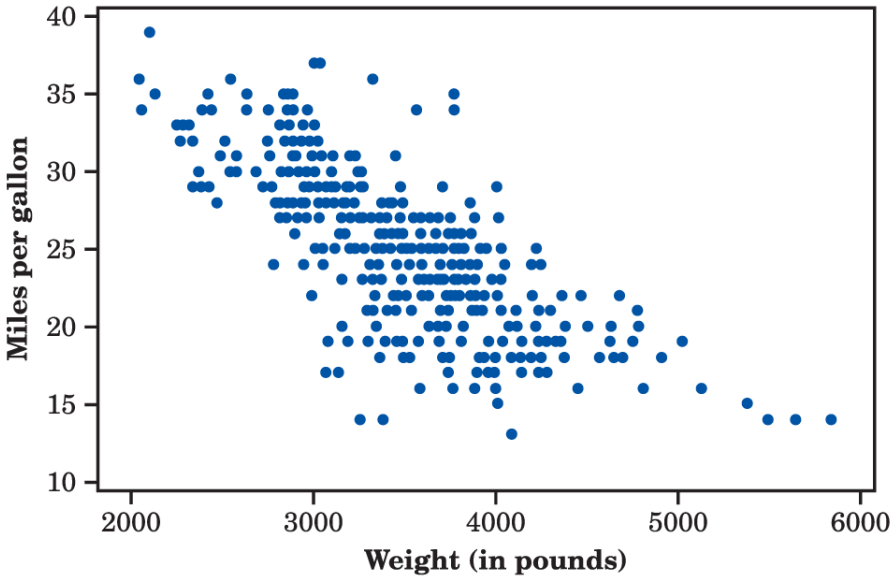


Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Scatterplots 3

The form on Figure 14.4 is a moderately strong linear relationship.

There is a negative association between gas mileage and weight.

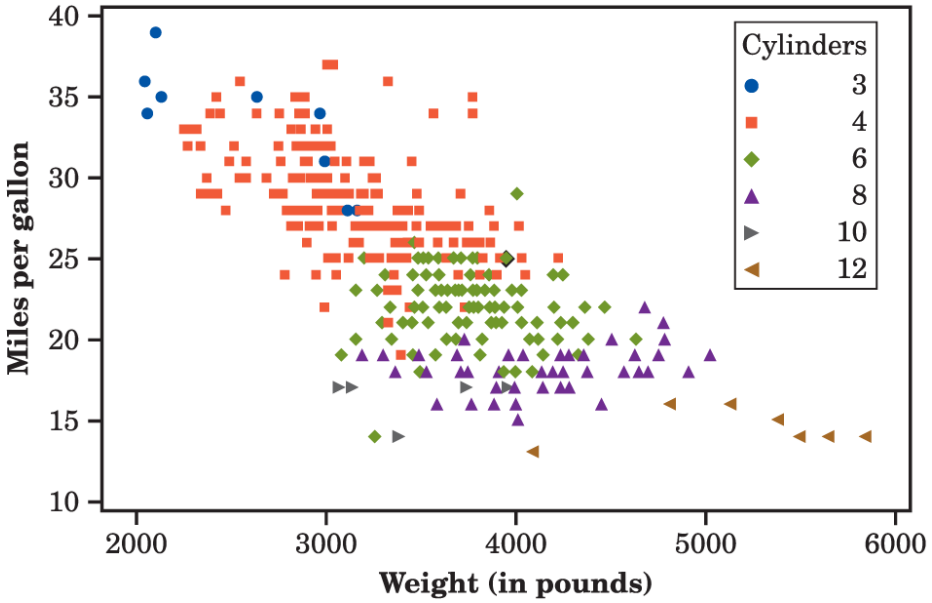


Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Multiple Variables

Multiple variables can be investigated at one time.

Figure 14.6 demonstrates what happens when number of cylinders is taken into account.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

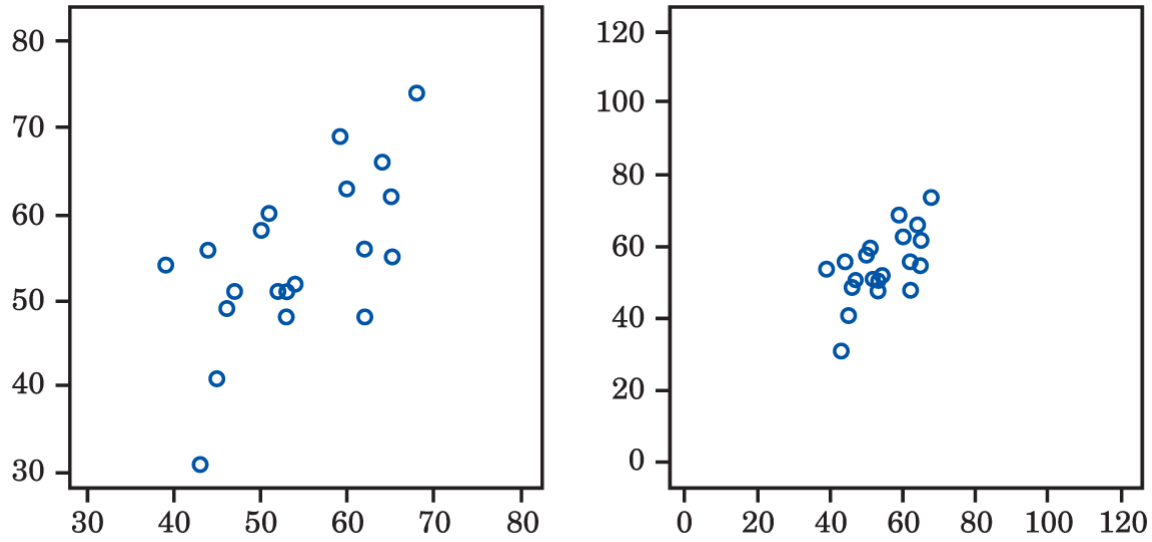
Correlation 1

A scatterplot displays the direction, form, and strength of the relationship between two variables.

Straight-line relations are particularly important because a straight line is a simple pattern that is quite common.

A straight-line relation is strong if the points lie close to a straight line, and weak if they are widely scattered about a line.

Correlation 2



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

The two scatterplots in Figure 14.8 depict the same data. The right-hand plot seems to show a stronger straight-line relationship. Our eyes can be fooled by changing the plotting scales or the amount of blank space around the cloud of points in a scatterplot. We need to follow our strategy for data analysis by using a numerical measure to supplement the graph.

Correlation 3

The **correlation** describes the direction and strength of a straight-line relationship between two quantitative variables. Correlation is usually written as *r*.

Calculating a correlation takes a bit of work. We usually calculate *r* using a calculator or software.

Understanding Correlation 1

Positive *r* indicates positive association between the variables, and negative *r* indicates negative association.

The correlation *r* always falls between −1 and 1. Values of *r* near 0 indicate a very weak straight-line relationship. The strength of the relationship increases as *r* moves away from 0 toward either −1 or 1. Values of *r* close to −1 or 1 indicate that the points lie close to a straight line. The extreme values $r = -1$ and $r = 1$ occur only when the points in a scatterplot lie exactly along a straight line.

Understanding Correlation 2

Because *r* uses the standard scores for the observations in its calculation, the correlation does not change when we change the units of measurement of *x*, *y*, or both.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

The correlation between two variables has no unit of measurement; it is just a number between −1 and 1.

Correlation ignores the distinction between explanatory and response variables. If we reverse our choice of which variable to call *x* and which to call *y*, the correlation does not change.

Understanding Correlation 3

Correlation measures the strength of only straight-line association between two variables. Correlation does not describe curved relationships between variables, no matter how strong they are.

The correlation is strongly affected by a few outlying observations. Use *r* with caution when outliers appear in the scatterplot.

Understanding Correlation 4

Correlation measures the *strength and direction* of a straight-line relationship between *two quantitative variables*.

Correlation is not a complete description of two-variable data, even when there is a straight-line relationship between the variables. You should give the means and standard deviations of both *x* and *y* along with the correlation. Because the formula for correlation uses the means and standard deviations, these measures are the proper choice to accompany a correlation.

Statistics in Summary 1

A **scatterplot** is a graph of the relationship between two quantitative variables. If you have an explanatory and a response variable, put the explanatory variable on the *x* (horizontal) axis of the scatterplot.

When you examine a scatterplot, look for the **direction, form, and strength** of the relationship and also for possible **outliers**.

If there is a clear direction, is it positive (the scatterplot slopes upward from left to right) or negative (the plot slopes downward)?

Statistics in Summary 2

- Is the form straight or curved? Are there clusters of observations? Is the relationship strong (a tight pattern in the plot) or weak (the points scatter widely)?
- The **correlation *r*** measures the direction and strength of a straight-line relationship between two quantitative variables.
- Correlation is a number between −1 and 1. The *r* shows whether the association is positive or negative. Its value gets closer to −1 or 1 as the points cluster more tightly around a straight line. The extreme values −1 and 1 occur only when the scatterplot shows a perfectly straight line.