into practice

Chapter 1
Statistical Modeling

Big Idea 1
Data consist of structure and variability.

> Essential Understanding 1*a*
> Mathematical models describe structure.
>
> Essential Understanding 1*b*
> Statistical models extend mathematical models by describing variability around the structure.
>
> Essential Understanding 1*c*
> Statistical models are evaluated by how well they describe data and whether they are useful.

To paraphrase the statistician George E. P. Box, all models are inaccurate, but some are useful (Box and Draper 1987, p. 424). The first big idea and its associated essential understandings in *Developing Essential Understanding of Statistics in Grades 9–12* (Peck, Gould, and Miller 2013) relate to the importance of structure and variability in understanding a set of data. Or, in the context of Box's observation, this big idea and its essential understandings address the question, "How inaccurate can a model be and still be useful?" These concepts also warn that statistical models differ from mathematical models in significant, even though sometimes subtle, ways. Statisticians need accurate models of data to make inferences, including predictions, about the population under study. The importance of accurate statistical models is reflected in the following high school standard from the Common Core State Standards for Mathematics (CCSSM; National Governors Association Center for Best Practices and Council of Chief State School Offices [NGA Center and CCSSO] 2010):

Understand and evaluate random processes underlying statistical experiments

2. Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation. (S-IC.2, p. 81)

What Constitutes a Statistical Model?

A statistical model can be as simple as pairing two variables by noticing that the value of one variable seems to be related to the value of the second variable. As an example, the table in figure 1.1 shows the number of oil changes per year for several cars, along with the associated cost of engine repairs. A model for this data might associate the annual cost of repairs for a car with the number of oil changes that the car has each year. In this case, the number of oil changes that the car has each year is the *predictor variable*, and the annual cost of repairs is the *response variable*.

Oil changes per year	X_1	3	5	2	3	1	4	6	4	3	2	0	10	7
Annual cost of repairs ($)	Y	300	300	500	400	700	400	100	250	450	650	600	0	150

Fig. 1.1. The number of oil changes per year and the annual cost of engine repairs for each of thirteen cars. From Illuminations, http://illuminations.nctm.org/Lesson.aspx?id=1189.

By contrast, someone else might believe that the annual cost of repairs for a car is more closely associated with the age of the car than with the number of oil changes that the car has each year. A model for this hypothesis might use the same response variable as the previous one (annual cost of repairs) but consider the car's age as the predictor variable. The table in figure 1.2 shows this pairing of variables.

Age of car (years)	X_2	4	3	5	5	6	4	3	3	6	8	3	2	5
Annual cost of repairs ($)	Y	300	300	500	400	700	400	100	250	450	650	600	0	150

Fig. 1.2. The age and annual repair cost for each of thirteen cars

A third person might believe that the best model would take *both* of these possible predictor variables into consideration, hypothesizing that a car's annual cost of repairs is associated with both its age and the number of oil changes that it has per year. A model for this hypothesis would have the same response variable as before (annual cost of repairs) but two predictor variables (number of oil changes per year and age of the car). The table in figure 1.3 displays this information for each car.

Oil changes per year	X_1	3	5	2	3	1	4	6	4	3	2	0	10	7
Age of car (years)	X_2	4	3	5	5	6	4	3	3	6	8	3	2	5
Annual cost of repairs ($)	Y	300	300	500	400	700	400	100	250	450	650	600	0	150

Fig. 1.3. Oil change, age, and annual repair cost for each of thirteen cars

Suppose that students were considering these three models. The question that they would want to address is, "Which of these models provides the best explanation of the difference in annual repair costs from car to car?" Knowing to ask this question and being able to answer it are at the heart of Essential Understanding 1c: "Statistical models are evaluated by how well they describe data and whether they are useful."

Inherent in understanding statistical models is a conceptual grounding in the idea of *variability*. The data in figures 1.1, 1.2, and 1.3 offer an opportunity to take a brief look at variability, which is the focus of Chapter 2. Figure 1.4 represents the data in the table in figure 1.1 in a scatterplot. A quick glance at the scatterplot would seem to support the idea that a relationship exists between the number of oil changes that a car has each year and its annual repair costs, and that a model describing this relationship could be constructed. However, your students might have difficulty verbalizing exactly what this means, since this relationship isn't a function in the mathematical sense.

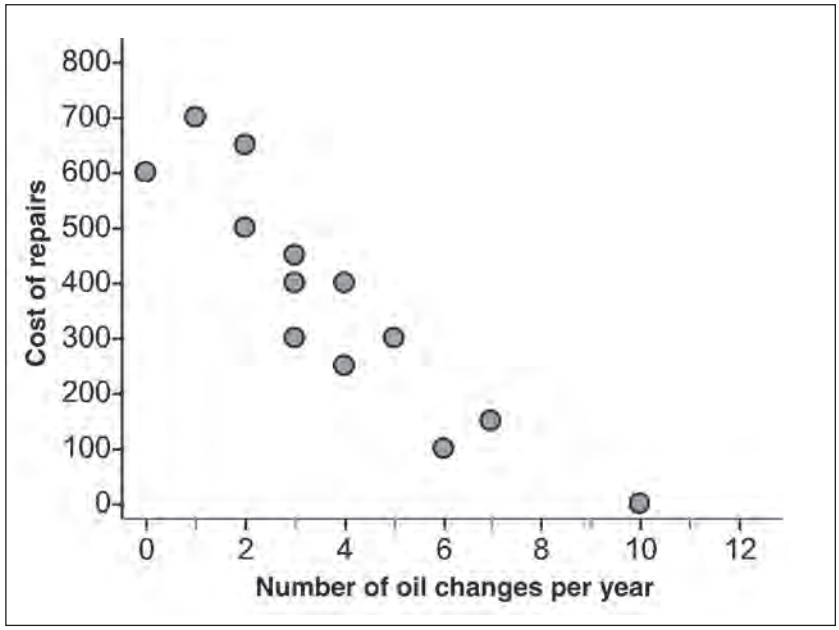

Fig. 1.4. A scatterplot of the data from the table in figure 1.1

In inspecting the table and the scatterplot, your students might note that some values of the predictor variable have unique corresponding values for the response variable—for example, (1, 700) and (10, 0). However, other values of the predictor variable have multiple corresponding values for the response variable—for instance, (3, 300), (3, 400), and (3, 450). So you might usefully ask your students to discuss precisely what it means to say that the annual cost of a car's repairs is associated with the number of oil changes that the car has per year. Reflect 1.1 invites you to consider how your students might approach this question. Pause and take time to think about your response to the questions posed in Reflect 1.1 (and, in their turn, all the subsequent questions for reflection that you encounter as you read the book).

Reflect 1.1

Suppose that your students had the table in figure 1.1 and the scatterplot in figure 1.4. How do you think they would respond if you asked, "What does it mean to say that the annual cost of a car's repairs is related to the number of oil changes that the car has per year?"

To interpret this model properly, students would need to realize that the data on these thirteen cars compose just a small sample of the data from all the cars that could have been included in the investigation. Furthermore, the underlying relationship under discussion concerns the *average* annual cost of repairs for all cars in the population from which the sample was drawn that have an equal number of oil changes per year. This is where the critically important idea of variability comes into play. For a given value X_1 of the predictor variable (number of oil changes per year), each observed value Y of the response variable (annual cost of repairs) can be expressed as $Y = \mu_1 + \varepsilon_1$, where μ_1 is the average annual cost of repairs for cars that have X_1 oil changes per year, and ε_1 measures how much Y, the annual cost of repairs for a selected car, varies from the average annual cost of repairs for all cars that have an equal number of oil changes per year. In general, the closer each ε_1 is to zero, the stronger the relationship is between X_1 and Y, and hence the more useful the model is in predicting annual cost of repairs.

A similar model could be used to describe the relationship between the variables in the table in figure 1.2. Figure 1.5 represents the data in that table in a scatterplot. In this case, the model can be expressed as $Y = \mu_2 + \varepsilon_2$, where μ_2 is the average annual cost of repairs for cars that are X_2 years old, and ε_2 is a measure of how much each Y_i differs from μ_2. Examine the questions in Reflect 1.2 to compare the usefulness of this model and the previous one for predicting a car's annual repair costs and to predict your students' choice.

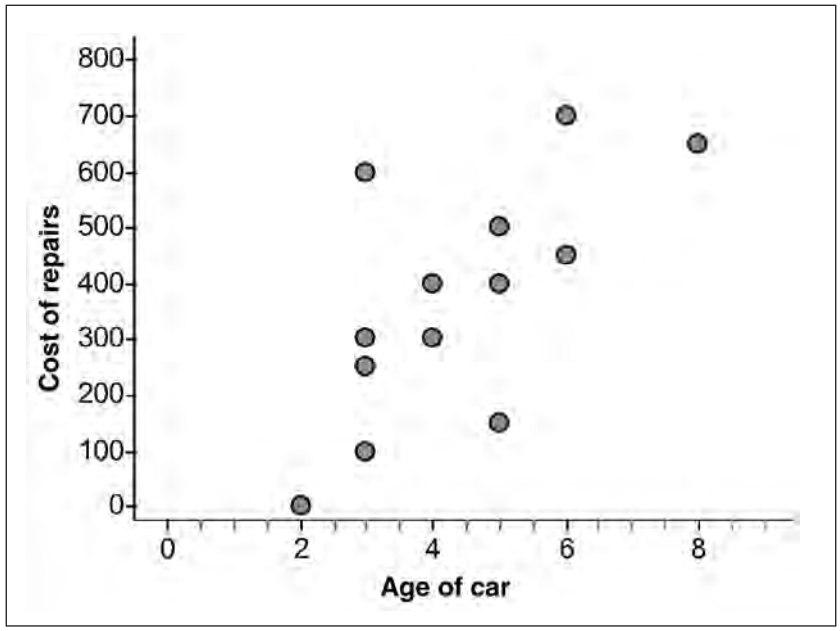

Fig. 1.5. A scatterplot of the data from the table in figure 1.2

Reflect 1.2

On the basis of the scatterplots in figures 1.4 and 1.5, which model do you think would be better for predicting a car's annual repair cost—the model using number of oil changes per year or the model using age of car?

How do you think your students would answer this question?

A linear relationship between X_1 and Y (or X_2 and Y) would be a specific, quantitative way of modeling the observed relationship in the scatterplot in figure 1.4 (or fig. 1.5). Least-squares techniques can be used to find the line of best fit for each of the scatterplots. The equation of the least-squares line in figure 1.6 is $\hat{Y} = 650.0 - 73.1X_1$, and the equation of the least-squares line in figure 1.7 is $\hat{Y} = -9.1 + 86.3X_2$. Students should be able to see that both of these models are "inaccurate" insofar as the fitted lines do not pass through each of the data points. However, the first model appears to be more "useful" since there is less variation in the points about the fitted regression line in figure 1.6 than there is in figure 1.7.

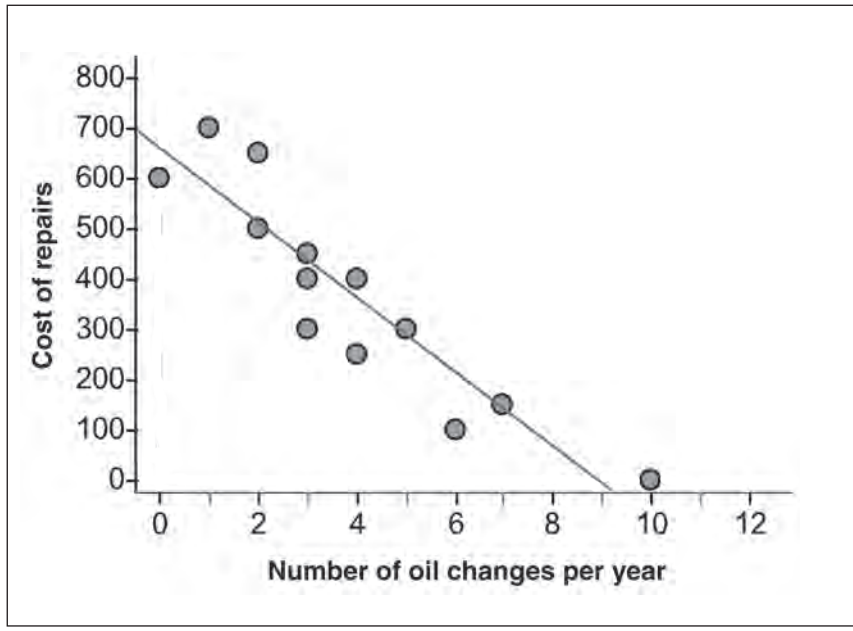

Fig. 1.6. The scatterplot from figure 1.4 along with the least-squares line

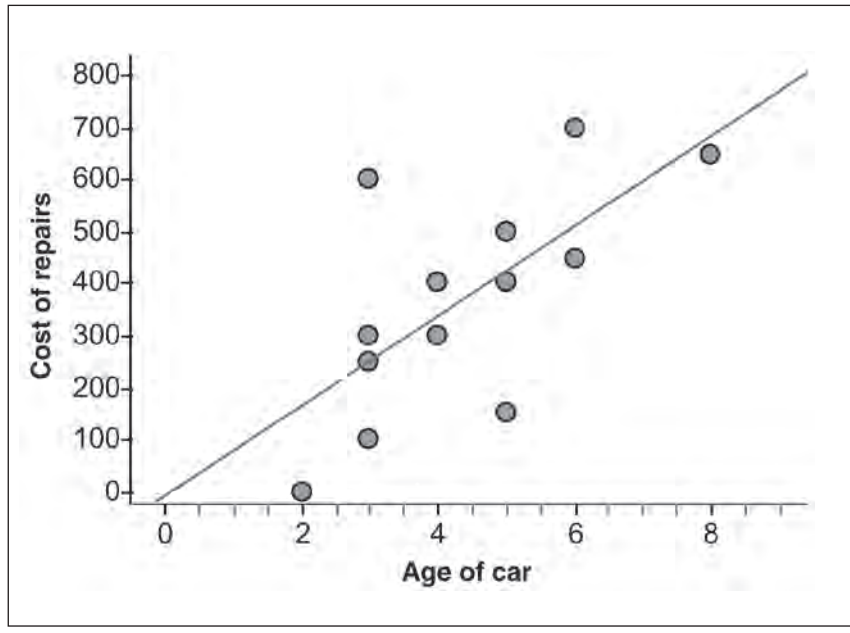

Fig. 1.7. The scatterplot from figure 1.5 along with the least-squares line

Finally, a linear model using both X_1 and X_2 can be built from the data in the table in figure 1.3; the least-squares line is $\hat{Y}$= 442.2 – 61.62X_1 + 37.42X_2. However, since the model has more than one predictor variable, this line cannot be displayed in a two-dimensional graph. Linear models with more than one predictor variable are beyond the scope of this book and will not be considered further.

Garfield and Ben-Zvi (2007) note that students tend to see a data set as individual values (each with its own characteristics) and not as an aggregate (a group with properties that may not be possessed by any individual member). Citing Hancock, Kaput, and Goldsmith (1992, p. 355), they go on to say, "To be able to think about the data as an aggregate, the aggregate must be constructed by the student" (p. 18).

In describing research efforts to help students make the conceptual leap between these two ways of viewing data, Ben-Zvi, Garfield, and Zieffler (2006) use the terms *local understanding* and *global understanding* of data and describe them as follows:

- Local understanding of data (or individual-based reasoning) involves focusing on an individual value or a few of them within a group of data (a particular entry in a table of data, a single point on a graph).

- Global understanding (or aggregate-based reasoning) refers to the ability to search for, recognize, describe, and explain general patterns in a set of data (change over time, trends) by naked-eye observation of distributions and by means of statistical parameters or techniques. (p. 470)

They further state, "Identifying patterns in a statistical graph depends on seeing the data set as a whole, taking into account the variability within the data, and integrating individual-based reasoning in some situations" (p. 472). They conjecture that the difficulty of making the transition from thinking about individual cases to aggregate-based reasoning may result from the differences between mathematical and statistical reasoning.

In working with a mathematical model for a relationship between variables, students come to expect that each point in a graph tells something about the deterministic relationship between variables, whereas in working with a statistical relationship, they must learn that the variation inherent in data means that being too focused on individual points' departures from a general trend can obscure their view of the general trend. Reflect 1.3 offers an opportunity to explore this difference with respect to the scatterplot in figure 1.8.

Reflect 1.3

Consider the data presented in figure 1.8 on U.S. coal production over the past six decades. Suppose that one student looks at the data from a local perspective and another looks at them from a global perspective.

How might the observations of the student with a local understanding differ from those of the student with a global view?

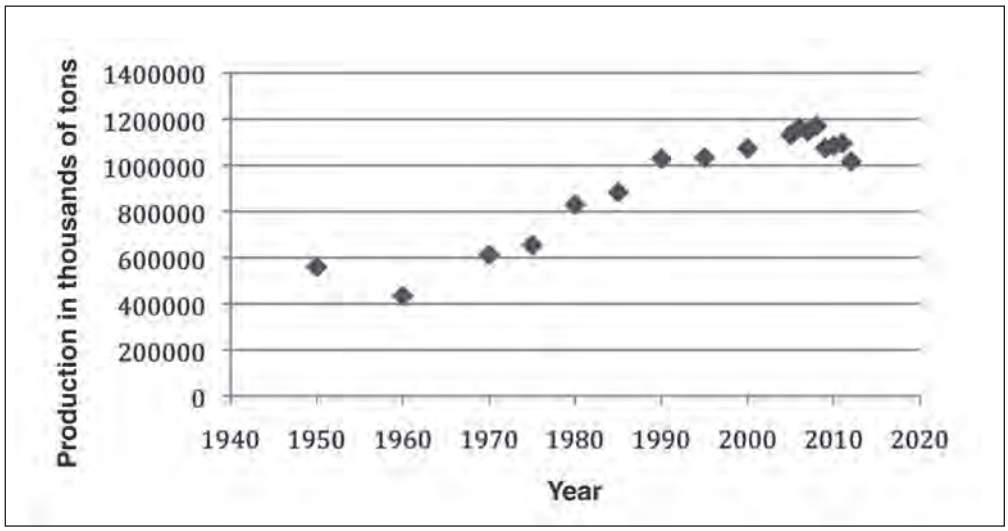

Fig. 1.8. U.S. coal production, 1950–2012

A student who exhibits only a local understanding of the data in figure 1.8 might notice that U.S. coal production was around 400 million tons in 1960 or that current annual production is at about 1 billion tons. A student with a global understanding of these same data might comment that U.S. coal production appears to have dipped between 1950 and 1960, generally increased from 1960 to about 2005, and has been decreasing since that time. Although a local understanding of data is often useful, the study of statistics is about *both* local *and* global understanding of data. As a result, students need to have a sufficient number of classroom experiences to allow them to develop a global understanding of data as well as a local understanding. Global understanding includes the ability to spot general trends and to see past the variation of individual points. However, students still must retain local understanding so that they can recognize when individual points depart from a general trend.