You can organize your work in any open-ended data analysis setting by following the four-step State, Plan, Solve, and Conclude process first introduced in Chapter 2. After we have mastered the extra background needed for statistical inference, this process will also guide practical work on inference later in the book.

PART I SUMMARY

Here are the most important skills you should have acquired from reading Chapters 1 to 6.

A. Data

1. Identify the individuals and variables in a set of data.
2. Identify each variable as categorical or quantitative. Identify the units in which each quantitative variable is measured.
3. Identify the explanatory and response variables in situations where one variable explains or influences another.

B. Displaying Distributions

1. Recognize when a pie chart can and cannot be used.
2. Make a bar graph of the distribution of a categorical variable, or in general to compare related quantities.
3. Interpret pie charts and bar graphs.
4. Make a histogram of the distribution of a quantitative variable.
5. Make a stemplot of the distribution of a small set of observations. Round leaves or split stems as needed to make an effective stemplot.
6. Make a time plot of a quantitative variable over time. Recognize patterns such as trends and cycles in time plots.

C. Describing Distributions (Quantitative Variable)

1. Look for the overall pattern and for major deviations from the pattern.
2. Assess from a histogram or stemplot whether the shape of a distribution is roughly symmetric, distinctly skewed, or neither. Assess whether the distribution has one or more major peaks.
3. Describe the overall pattern by giving numerical measures of center and spread in addition to a verbal description of shape.
4. Decide which measures of center and spread are more appropriate: the mean and standard deviation (especially for symmetric distributions) or the five-number summary (especially for skewed distributions).
5. Recognize outliers and give plausible explanations for them.

D. Numerical Summaries of Distributions

1. Find the median M and the quartiles Q_1 and Q_3 for a set of observations.
2. Find the five-number summary and draw a boxplot; assess center, spread, symmetry, and skewness from a boxplot.

3. Find the mean $\overline{x}$ and the standard deviation s for a set of observations.
4. Understand that the median is more resistant than the mean. Recognize that skewness in a distribution moves the mean away from the median toward the long tail.
5. Know the basic properties of the standard deviation: $s \geq 0$ always; $s = 0$ only when all observations are identical and increases as the spread increases; s has the same units as the original measurements; s is pulled strongly up by outliers or skewness.

E. Density Curves and Normal Distributions

1. Know that areas under a density curve represent proportions of all observations and that the total area under a density curve is 1.
2. Approximately locate the median (equal-areas point) and the mean (balance point) on a density curve.
3. Know that the mean and median both lie at the center of a symmetric density curve and that the mean moves farther toward the long tail of a skewed curve.
4. Recognize the shape of Normal curves and estimate by eye both the mean and standard deviation from such a curve.
5. Use the 68–95–99.7 rule and symmetry to state what percent of the observations from a Normal distribution fall between two points when both points lie at the mean or one, two, or three standard deviations on either side of the mean.
6. Find the standardized value (z-score) of an observation. Interpret z-scores and understand that any Normal distribution becomes standard Normal $N(0, 1)$ when standardized.
7. Given that a variable has a Normal distribution with a stated mean μ and standard deviation σ, calculate the proportion of values above a stated number, below a stated number, or between two stated numbers.
8. Given that a variable has a Normal distribution with a stated mean μ and standard deviation σ, calculate the point having a stated proportion of all values above it or below it.

F. Scatterplots and Correlation

1. Make a scatterplot to display the relationship between two quantitative variables measured on the same subjects. Place the explanatory variable (if any) on the horizontal scale of the plot.
2. Add a categorical variable to a scatterplot by using a different plotting symbol or color.
3. Describe the direction, form, and strength of the overall pattern of a scatterplot. In particular, recognize positive or negative association and linear (straight-line) patterns. Recognize outliers in a scatterplot.
4. Judge whether it is appropriate to use correlation to describe the relationship between two quantitative variables. Find the correlation r.

5. Know the basic properties of correlation: r measures the direction and strength of only straight-line relationships; r is always a number between -1 and 1; $r = \pm 1$ only for perfect straight-line relationships; r moves away from 0 toward ± 1 as the straight-line relationship gets stronger.

G. Regression Lines

1. Understand that regression requires an explanatory variable and a response variable. Use a calculator or software to find the least-squares regression line of a response variable y on an explanatory variable x from data.
2. Explain what the slope b and the intercept a mean in the equation $\hat{y} = a + bx$ of a regression line.
3. Draw a graph of a regression line when you are given its equation.
4. Use a regression line to predict y for a given x. Recognize extrapolation and be aware of its dangers.
5. Find the slope and intercept of the least-squares regression line from the means and standard deviations of x and y and their correlation.
6. Use r^2, the square of the correlation, to describe how much of the variation in one variable can be accounted for by a straight-line relationship with another variable.
7. Recognize outliers and potentially influential observations from a scatterplot with the regression line drawn on it.
8. Calculate the residuals and plot them against the explanatory variable x. Recognize that a residual plot magnifies the pattern of the scatterplot of y versus x.

H. Cautions about Correlation and Regression

1. Understand that both r and the least-squares regression line can be strongly influenced by a few extreme observations.
2. Recognize possible lurking variables that may explain the observed association between two variables x and y.
3. Understand that even a strong correlation does not mean that there is a cause-and-effect relationship between x and y.
4. Give plausible explanations for an observed association between two variables: direct cause and effect, the influence of lurking variables, or both.

I. Categorical Data (Optional)

1. From a two-way table of counts, find the marginal distributions of both variables by obtaining the row sums and column sums.
2. Express any distribution in percents by dividing the category counts by their total.
3. Describe the relationship between two categorical variables by computing and comparing percents. Often this involves comparing the conditional distributions of one variable for the different categories of the other variable.
4. Recognize Simpson's paradox and be able to explain it.

Driving in Canada

Canada is a civilized and restrained nation, at least in the eyes of Americans. A survey sponsored by the Canada Safety Council suggests that driving in Canada may be more adventurous than expected. Of the Canadian drivers surveyed, 88% admitted to aggressive driving in the past year, and 76% said that sleep-deprived drivers were common on Canadian roads. What really alarms us is the name of the survey: the Nerves of Steel Aggressive Driving Study.