GLOBAL EDITION

Business Statistics

A First Course

8E

David M. Levine
Kathryn A. Szabat
David F. Stephan

A Roadmap for Selecting a Statistical Method

Data Analysis Task	For Numerical Variables	For Categorical Variables
Describing a group or several groups	Ordered array, stem-and-leaf display, frequency distribution, relative frequency distribution, percentage distribution, cumulative percentage distribution, histogram, polygon, cumulative percentage polygon **(Sections 2.2, 2.4)** Mean, median, mode, geometric mean, quartiles, range, interquartile range, standard deviation, variance, coefficient of variation, skewness, kurtosis, boxplot, normal probability plot **(Sections 3.1, 3.2, 3.3, 6.3)** Dashboards **(Section 14.2)**	Summary table, bar chart, pie chart, doughnut chart, Pareto chart **(Sections 2.1 and 2.3)**
Inference about one group	Confidence interval estimate of the mean **(Sections 8.1 and 8.2)** *t* test for the mean **(Section 9.2)**	Confidence interval estimate of the proportion **(Section 8.3)** *Z* test for the proportion **(Section 9.4)**
Comparing two groups	Tests for the difference in the means of two independent populations **(Section 10.1)** Paired *t* test **(Section 10.2)** *F* test for the difference between two variances **(Section 10.4)**	*Z* test for the difference between two proportions **(Section 10.3)** Chi-square test for the difference between two proportions **(Section 12.1)**
Comparing more than two groups	One-way analysis of variance for comparing several means **(Section 11.1)**	Chi-square test for differences among more than two proportions **(Section 12.2)**
Analyzing the relationship between two variables	Scatter plot, time series plot **(Section 2.5)** Covariance, coefficient of correlation **(Section 3.5)** Simple linear regression **(Chapter 13)** *t* test of correlation **(Section 13.7)** Sparklines **(Section 2.7)**	Contingency table, side-by-side bar chart, PivotTables **(Sections 2.1, 2.3, 2.6)** Chi-square test of independence **(Section 12.3)**
Analyzing the relationship between two or more variables	Colored scatter plots, bubble chart, treemap **(Section 2.7)** Multiple regression **(Chapters 14)** Dynamic bubble charts **(Section 14.2)** Regression trees **(Section 14.3)** Cluster analysis **(Section 14.5)** Multidimensional scaling **(Section 14.6)**	Multidimensional contingency tables **(Section 2.6)** Drilldown and slicers **(Section 2.7)** Classification trees **(Section 14.4)** Multiple correspondence analysis **(Section 14.6)**

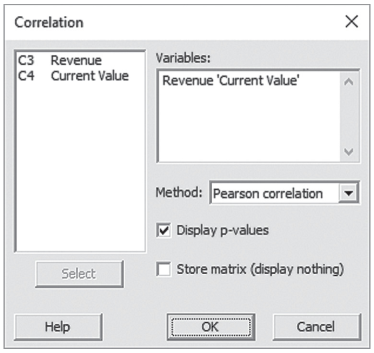
In the table of numbers produced, the covariance is the number that appears in the cell position that is the intersection of the two variables (the lower-left cell).

The Coefficient of Correlation

Use **Correlation**.

For example, to compute the coefficient of correlation for Example 3.17 on page 189, open the **NBAValues worksheet**. Select **Stat ➔ Basic Statistics ➔ Correlation**. In the Correlation dialog box (shown at right):

1. Double-click **C3 Revenue** in the variables list to add **Revenue** to the **Variables** box.
2. Double-click **C4 Current Value** in the variables list to add **'Current Value'** to the **Variables** box.
3. Select **Pearson correlation** from the **Method** pull-down list.
4. Check **Display p-values**.
5. Click **OK**.



CHAPTER 3

TABLEAU GUIDE

TG3.3 EXPLORING NUMERICAL VARIABLES

The Five-Number Summary and the Boxplot

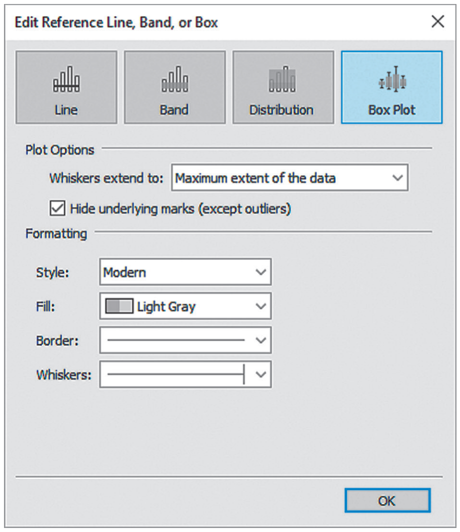
Use **box-and-whisker plots**.

For example, to construct the five-number summary boxplots of the three-year return percentage variable for the growth and value funds, similar to Figure 3.8 on page 180, open to the **Retirement Funds For Tableau Excel workbook**. In a new Tableau worksheet:

1. Right-click **3YrReturn** in the Dimensions list and select **Convert to Measure** in the shortcut menu.
2. Drag **3YrReturn** from the Measures list and drop it in the **Rows** area.
3. Drag **Fund Number** from the Dimensions list and drop it in the **Rows** area.
4. If the ShowMe tab is not visible, click **ShowMe**.
5. In the ShowMe tab, click **box-and-whisker plots**.
6. Drag **Fund Type** from the Dimensions list and drop it in the **Rows** area.
7. Right-click over one of the boxes and select **Edit** from the shortcut menu.

In the Edit Reference Line, Band, or Box dialog box (shown in the next column):

8. Select **Maximum extent of the data** from the **Plot Options** pull-down list.
9. Check **Hide underlying marks (except outliers)**.
10. Click **OK**.



11. Back in the main Tableau window, select **Analysis ➔ Swap Rows and Columns**.

Boxplots change from a vertical to horizontal layout. Moving the mouse pointer over each boxplot, displays the values of five-number summary of the boxplot. Tableau labels the first quartile, Q_1, the Lower Hinge, and labels the third quartile, Q_3, the Upper Hinge. (The lower and upper hinges are equivalent to these quartiles by the method this book uses to calculate Q_1 and Q_3.)

To construct a boxplot for a numerical variable, the variable must be a measure and not a dimension, the reason for step 1.

4 Basic Probability

CONTENTS

OBJECTIVES

- Understand basic probability concepts
- Understand conditional probability
- Use Bayes' theorem to revise probabilities
- Apply counting rules

▼USING **STATISTICS**
Possibilities at M&R Electronics World

As the marketing manager for M&R Electronics World, you are analyzing the results of an intent-to-purchase study. The heads of 1,000 households were asked about their intentions to purchase a large TV (screen size of at least 60 inches, measured diagonally) sometime during the next 12 months. As a follow-up, you plan to survey the same people 12 months later to see whether they purchased a large TV. For households that did purchase a large TV, you would like to know whether the television they purchased had a faster refresh rate (120 Hz or higher) or a standard refresh rate (60 Hz), whether they also purchased a streaming media player in the past 12 months, and whether they were satisfied with their purchase of the large TV.

You plan to use the results of this survey to form a new marketing strategy that will enhance sales and better target those households likely to purchase multiple or more expensive products. What questions can you ask in this survey? How can you express the relationships among the various intent-to-purchase responses of individual households?

The principles of probability help bridge the worlds of descriptive statistics and inferential statistics. Probability principles are the foundation for the probability distribution, the concept of mathematical expectation, and the binomial and Poisson distributions. In this chapter, you will learn to apply probability to intent-to-purchase survey responses to answer purchase behavior questions such as:

- What is the probability that a household is planning to purchase a large TV in the next year?
- What is the probability that a household will actually purchase a large TV?
- What is the probability that a household is planning to purchase a large TV and actually purchases the television?
- Given that the household is planning to purchase a large TV, what is the probability that the purchase is made?
- Does knowledge of whether a household *plans* to purchase a large TV change the likelihood of predicting whether the household *will* purchase a large TV?
- What is the probability that a household that purchases a large TV will purchase a television with a faster refresh rate?
- What is the probability that a household that purchases a large TV with a faster refresh rate will also purchase a streaming media player?
- What is the probability that a household that purchases a large TV will be satisfied with the purchase?

With answers to questions such as these, you can begin to form a marketing strategy. You can consider whether to target households that have indicated an intent to purchase or to focus on selling televisions that have faster refresh rates or both. You can also explore whether households that purchase large TVs with faster refresh rates can be easily persuaded to also purchase streaming media players.

4.1 Basic Probability Concepts

In everyday usage, *probability*, according to the Oxford English Dictionary, indicates the extent to which something is likely to occur or exist but can also mean the most likely cause of something. If storm clouds form, the wind shifts, and the barometric pressure drops, the probability of rain coming soon increases (first meaning). If one observes people entering an office building with wet clothes or otherwise drenched, there is a strong probability that it is currently raining outside (second meaning).

In statistics, **probability** is a numerical value that expresses the ratio between the value sought and the set of all possible values that could occur. A six-sided die has faces for 1, 2, 3, 4, 5, and 6. Therefore, for one roll of a *fair* six-sided die, the set of all possible values are the values 1 through 6. If the value sought is "a value greater than 4," then the values 5 or 6 would be sought. One would say the probability of this *event* is 2 outcomes divided by 6 outcomes or 1/3.

Consider tossing a fair coin heads or tails two times. What is the probability of tossing two tails? The set of possible values for tossing a fair coin twice are HH, TT, HT, TH. Therefore, the probability of tossing two tails is 1/4 because only one value (TT) matches what is being sought and there are 4 values in the set of all possible values.

Events and Sample Spaces

When discussing probability, one formally uses **outcomes** in place of *values* and calls the set of all possible outcomes the **sample space**. **Events** are subsets of the sample space, the set of all outcomes that produce a specific result. For tossing a fair coin twice, the event "toss at least 1 head" is the subset of outcomes HH, HT, and TH, and the event "toss two tails" is the subset TT. Both of these events are also examples of a **joint event**, an event that has two or more characteristics. In contrast, a **simple event** has only one characteristic, an outcome that cannot be further subdivided. The event "rolling a value greater 4" in the first example results in the subset of outcomes 5 and 6 and is an example of a simple event because "5" and "6" represent one characteristic and cannot be further divided.

student TIP

Events are represented by letters of the alphabet.

The **complement** of an event *A*, noted by the symbol A', is the subset of outcomes that are not part of the event. For tossing a fair coin twice, the complement of the event "toss at least 1 head" is the subset TT, while the complement of the event "toss two tails" is HH, HT, and TH.

A set of events are **mutually exclusive** if they cannot occur at the same. The events "roll a value greater than 4" and "roll a value less than 3" are mutually exclusive when rolling one fair die. However, the events "roll a value greater than 4" and "roll a value greater than 5" are not because both share the outcome of rolling a 6.

A set of events are **collectively exhaustive** if one of the events must occur. For rolling a fair six-sided die, the events "roll a value 3 or less" and "roll a value 4 or more" are collectively exhaustive because these two subsets include all possible outcomes in the sample space. However, the set of events "roll a value 3 or less" and "roll a value greater than 4" is not because this set does not include the outcome of rolling a 4.

student TIP

By definition, *an event and its complement* are always both mutually exclusive and collectively exhaustive.

Not all sets of collectively exhaustive events are mutually exclusive. For rolling a fair six-sided die, the set of events "roll a value 3 or less," "roll an even numbered value," and "roll a value greater than 4" is collectively exhaustive but is not mutually exclusive as, for example, "a value 3 or less" and "an even numbered value" could *both* occur if a 2 is rolled.

student TIP

A probability cannot be negative or greater than 1.

Certain and *impossible* events represent special cases. A **certain event** is an event that is sure to occur such as "roll a value greater than 0" for rolling one fair die. Because the subset of outcomes for a certain event is the entire set of outcomes in the sample, a certain event has a probability of 1. An **impossible event** is an event that has no chance of occurring, such as "roll a value greater than 6" for rolling one fair die. Because the subset of outcomes for an impossible event is empty—contains no outcomes—an impossible event has a probability of 0.

Types of Probability

The concepts and vocabulary related to events and sample spaces are helpful to understanding how to calculate probabilities. Also affecting such calculations is the type of probability being used: *a priori*, empirical, or subjective.

In ***a priori* probability**, the probability of an occurrence is based on having prior knowledge of the outcomes that can occur. Consider a standard deck of cards that has 26 red cards and 26 black cards. The probability of selecting a black card is $26/52 = 0.50$ because there are 26 black cards and 52 total cards. What does this probability mean? If each card is replaced after it is selected, this probability does not mean that 1 out of the next 2 cards selected will be black. One cannot say for certain what will happen on the next several selections. However, one can say that in the long run, if this selection process is continually repeated, the proportion of black cards selected will approach 0.50. Example 4.1 shows another example of computing an *a priori* probability.

EXAMPLE 4.1

Finding *A Priori* Probabilities

A standard six-sided die has six faces. Each face of the die contains either one, two, three, four, five, or six dots. If you roll a die, what is the probability that you will get a face with five dots?

SOLUTION Each face is equally likely to occur. Because there are six faces, the probability of getting a face with five dots is 1/6.

The preceding examples use the *a priori* probability approach because the number of ways the event occurs and the total number of possible outcomes are known from the composition of the deck of cards or the faces of the die.

In the **empirical probability** approach, the probabilities are based on observed data, not on prior knowledge of how the outcomes can occur. Surveys are often used to generate empirical probabilities. Examples of this type of probability are the proportion of individuals in the M&R Electronics World scenario who actually purchase a large TV, the proportion of registered voters who prefer a certain political candidate, and the proportion of students who have part-time jobs. For example, if one conducts a survey of students, and 60% state that they have part-time jobs, then there is a 0.60 probability that an individual student has a part-time job.

The third approach to probability, **subjective probability**, differs from the other two approaches because subjective probability differs from person to person. For example, the development team for a new product may assign a probability of 0.60 to the chance of success for the product, while the president of the company may be less optimistic and assign a probability of 0.30. The assignment of subjective probabilities to various outcomes is usually based on a combination of an individual's past experience, personal opinion, and analysis of a particular situation. Subjective probability is especially useful in making decisions in situations in which one cannot use *a priori* probability or empirical probability.

Summarizing Sample Spaces

Sample spaces can be presented in tabular form using contingency tables (see Section 2.1) or visualized using Venn diagrams. Table 4.1 in Example 4.2 summarizes a sample space as a contingency table. When used for probability, each cell in a contingency table represents one joint *event*, analogous to the one joint *response* when these tables are used to summarize categorical variables. For example, 200 of the respondents correspond to the joint event "planned to purchase a large TV and subsequently did purchase the large TV."

EXAMPLE 4.2

Events and Sample Spaces

The M&R Electronics World scenario on page 206 concerns analyzing the results of an intent-to-purchase study. Table 4.1 presents the results of the sample of 1,000 households surveyed in terms of purchase behavior for large TVs.

TABLE 4.1
Purchase Behavior for Large TVs

PLANNED TO PURCHASE	ACTUALLY PURCHASED		
	Yes	**No**	**Total**
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

What is the sample space? Give examples of simple events and joint events.

SOLUTION The sample space consists of the 1,000 respondents. Simple events are "planned to purchase," "did not plan to purchase," "purchased," and "did not purchase." The complement of the event "planned to purchase" is "did not plan to purchase." The event "planned to purchase and actually purchased" is a joint event because in this joint event, the respondent must plan to purchase the television *and* actually purchase it.

Venn diagrams visualize a sample space. This diagram represents the various events as "unions" and "intersections" of circles. Figure 4.1 presents a typical Venn diagram for a two-variable situation, with each variable having only two events (A and A', B and B'). The circle on the left (the red one) represents all events that are part of A.

FIGURE 4.1
Venn diagram for events A and B

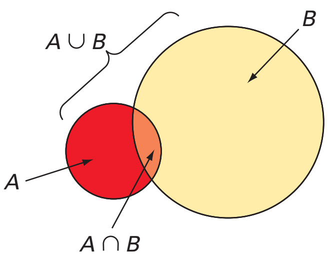
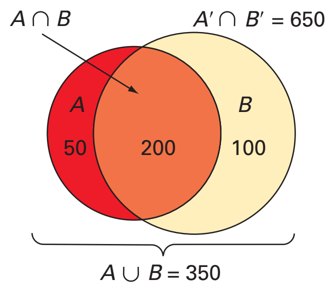


FIGURE 4.2
Venn diagram for the M&R Electronics World example



The circle on the right (the yellow one) represents all events that are part of B. The area contained within circle A and circle B (center area) is the intersection of A and B (written as $A \cap B$) because it is part of A and also part of B. The total area of the two circles is the union of A and B (written as $A \cup B$) and contains all outcomes that are just part of event A, just part of event B, or part of both A and B. The area in the diagram outside of $A \cup B$ contains outcomes that are neither part of A nor part of B.

To develop a Venn diagram, one must first define A and B. One can define either event as A or B but must be consistent in their definitions when evaluating the various events. For the Example 4.2 large TV example, one can define the events:

$$A = \text{planned to purchase} \qquad B = \text{actually purchased}$$

$$A' = \text{did not plan to purchase} \qquad B' = \text{did not actually purchase}$$

In drawing the Venn diagram for this problem (see Figure 4.2 on page 209), first determine the value of the intersection of A and B so that the sample space can be divided into its parts. $A \cap B$ consists of all 200 households who planned to purchase and actually purchased a large TV. The remainder of event A (planned to purchase) consists of the 50 households who planned to purchase a large TV but did not actually purchase one. The remainder of event B (actually purchased) consists of the 100 households who did not plan to purchase a large TV but actually purchased one. The remaining 650 households represent those who neither planned to purchase nor actually purchased a large TV.

Simple Probability

Simple probability is the probability of occurrence of a simple event A, $P(A)$, in which each outcome is *equally likely* to occur. Equation (4.1) defines the probability of occurrence for simple probability.

PROBABILITY OF OCCURRENCE

$$\text{Probability of occurrence} = \frac{X}{T} \qquad \textbf{(4.1)}$$

where

X = number of outcomes in which the event occurs
T = total number of possible outcomes

Equation 4.1 represents what some people wrongly think *is* the probability of occurrence for *all* probability problems. (Not all probability problems can be solved by Equation 4.1 as later examples in this chapter illustrate.) In the M&R Electronics World scenario, the collected survey data represent an example of empirical probability. Therefore, one can use Equation (4.1) to determine answers to questions that can be expressed as a simple probability.

For example, one question asked respondents if they planned to purchase a large TV. Using the responses to this question, how can one determine the probability of selecting a household that planned to purchase a large TV? From the Table 4.1 contingency table, determine the value of X as 250, the total of the Planned-to-Purchase Yes row and determine the value of T as 1,000, the overall total of respondents located in the lower right corner cell of the table. Using Equation (4.1) and Table 4.1 or Figure 4.2:

$$\text{Probability of occurrence} = \frac{X}{T}$$

$$P(\text{Planned to purchase}) = \frac{\text{Number who planned to purchase}}{\text{Total number of households}}$$

$$= \frac{250}{1{,}000} = 0.25$$

Thus, there is a 0.25 (or 25%) chance that a household planned to purchase a large TV.