IN THIS CHAPTER

- Looking at the key properties of data
- Understanding probability's role in business
- Sampling distributions
- Drawing conclusions based on results

Chapter 1

The Art and Science of Business Statistics

Statistical analysis is widely used in all business disciplines. For example, marketing researchers analyze consumer spending patterns to properly plan new advertising campaigns. Organizations use management consulting to determine how efficiently resources are being used. Manufacturers use quality control methods to ensure the consistency of the products they are producing. These types of business applications and many others are heavily based on statistical analysis.

Financial institutions use statistics for a wide variety of applications. For example, a pension fund may use statistics to identify the types of securities that it should hold in its investment portfolio. A hedge fund may use statistics to identify profitable trading opportunities. An investment bank may forecast the future state of the economy to determine which new assets it should hold in its own portfolio.

Whereas statistics is a quantitative discipline, the ultimate objective of statistical analysis is to explain real-world events. This means that in addition to the rigorous application of statistical methods, there is always a great deal of room for judgment. As a result, you can think of statistical analysis as both a science and an art; the art comes from choosing the appropriate statistical technique for a given situation and correctly interpreting the results.

In this chapter, I provide a brief introduction to the concepts that are covered throughout the book. I introduce several important techniques that help you to measure and analyze the statistical properties of real-world variables, such as stock prices, interest rates, corporate profits, and so on.

Representing the Key Properties of Data

The word *data* refers to a collection of *quantitative* (numerical) or *qualitative* (non-numerical) values. Quantitative data may consist of prices, profits, sales, or any variable that can be measured on a numerical scale. Qualitative data may consist of colors, brand names, geographic locations, and so on. Most of the data encountered in business applications are quantitative.


TECHNICAL STUFF

The word *data* is actually the plural of *datum;* datum refers to a single value, while data refers to a collection of values.

You can analyze data with graphical techniques or numerical measures. I explore both options in the following sections.

Analyzing data with graphs

Graphs are a visual representation of a data set, making it easy to see patterns and other details. Deciding which type of graph to use depends on the type of data you're trying to analyze. Here are some of the more common types of graphs used in business statistics:

- **Histograms:** A histogram shows the distribution of data among different intervals or categories, using a series of vertical bars.
- **Line graphs:** A line graph shows how a variable changes over time.
- **Pie charts:** A pie chart shows how data is distributed between different categories, illustrated as a series of slices taken from a pie.
- **Scatter plots (scatter diagrams):** A scatter plot shows the relationship between two variables as a series of points. The pattern of the points indicates how closely related the two variables are.

Histograms

You can use a histogram with either quantitative or qualitative data. It's designed to show how a variable is distributed among different categories. For example,

suppose a marketing firm surveys 100 consumers to determine their favorite color. The responses are

Red:	23
Blue:	44
Yellow:	12
Green:	21

The results can be illustrated with a histogram, with each color in a single category. The heights of the bars indicate the number of responses for each color, making it easy to see which colors are the most popular (see Figure 1-1).

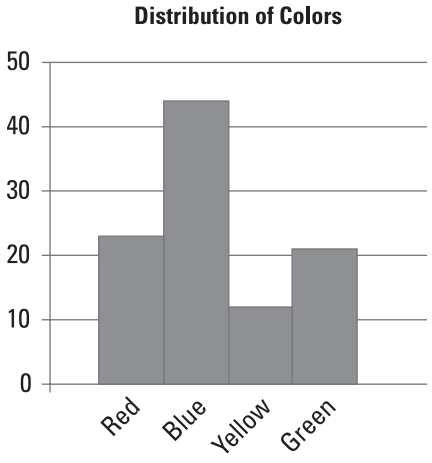

FIGURE 1-1: A histogram for preferred colors.

Based on the histogram, you can see at a glance that blue is the most popular choice, while yellow is the least popular choice.

Line graphs

You can use a line graph with quantitative data. It shows the values of a variable over a given interval of time. For example, Figure 1-2 shows the daily price of gold between August 1, 2023 and September 29, 2023.

With a line graph, it's easy to see trends or patterns in a data set. These types of graphs may be used by investors to identify which assets are likely to rise in the future based on their past performance.

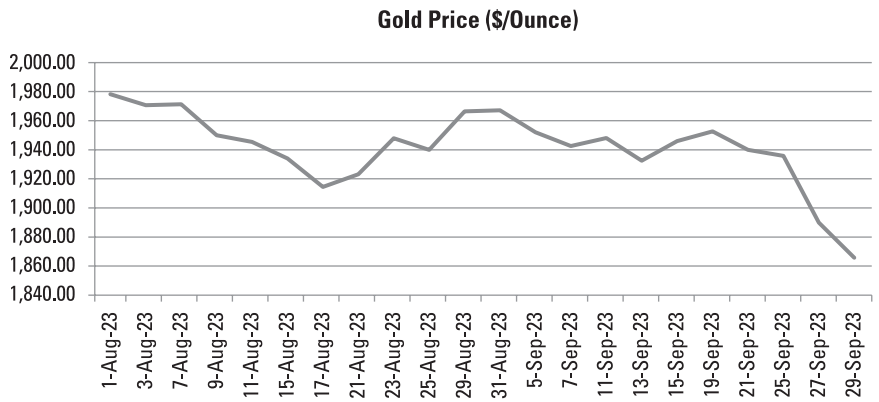

FIGURE 1-2: A line graph of gold prices.

Pie charts

Use a pie chart with quantitative or qualitative data to show the distribution of the data among different categories. For example, suppose that a chain of coffee shops wants to analyze its sales by coffee style. The styles that the chain sells are French Roast, Breakfast Blend, Brazilian Rainforest, Jamaica Blue Mountain, and Espresso. Figure 1-3 shows the proportion of sales for each style.

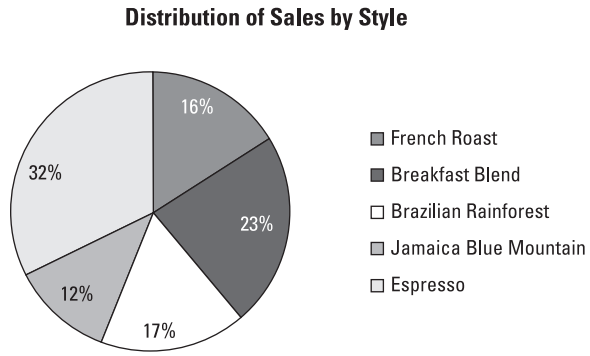

FIGURE 1-3: A pie chart for coffee sales.

The chart shows that Espresso is the chain's best-selling style, while Jamaica Blue Mountain accounts for the smallest percentage of the chain's sales.

Scatter plots

A scatter plot is designed to show the relationship between two quantitative variables. For example, Figure 1-4 shows the relationship between a corporation's sales and profits over the past 20 years.

Each point on the scatter plot represents profit and sales for a single year. The pattern of the points shows that higher levels of sales tend to be matched by higher levels of profits, and vice versa. This is called a positive relationship between the two variables.

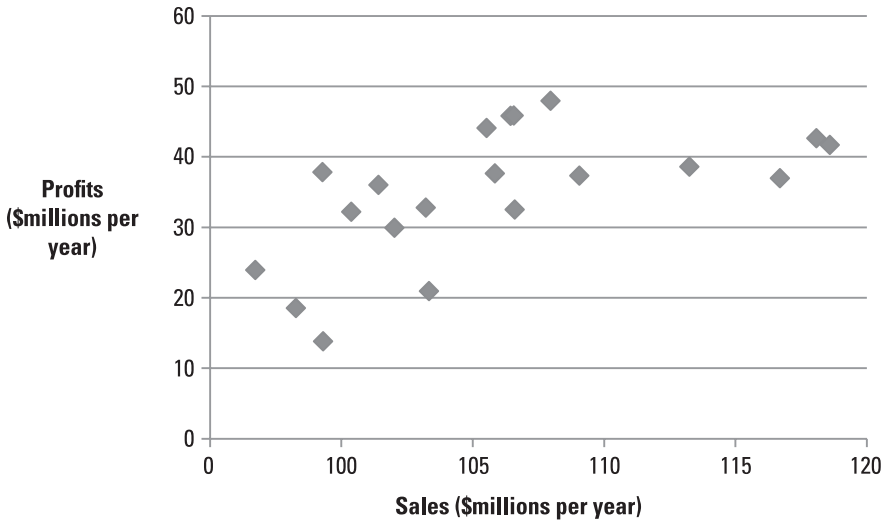

FIGURE 1-4: A scatter plot showing sales and profits.

Defining properties and relationships with numerical measures

A *numerical measure* is a value that describes a key property of a data set. For example, to determine whether the residents of one city tend to be older than the residents in another city, you can compute and compare the average or *mean* age of the residents of each city. Some of the most important properties of interest in a data set are the *center* of the data and the *spread* among the observations.

Finding the center of the data

To identify the center of a data set, you use measures that are known as *measures of central tendency;* the most important of these are the mean, median, and mode.

The *mean* represents the average value in a data set, while the *median* represents the midpoint. The median is a value that separates the data into two halves; half of the elements in the data set are *less than or equal to* the median, and the remaining half are *greater than or equal to* the median. The *mode* is the most commonly occurring value in the data set.

The mean is the most widely used measure of central tendency, but it can give deceptive results if the data contain any unusually large or small values, known as *outliers*. In this case, the median provides a more representative measure of the center of the data. For example, median household income is usually reported by government agencies instead of mean household income. This is because mean household income is inflated by the presence of a small number of extremely wealthy households. As a result, median household income is thought to be a better measure of how standards of living are changing over time.

The mode can be used for either quantitative or qualitative data. For example, it may be used to determine the most common number of years of education among the employees of a firm. It may also be used to determine the most popular flavor sold by a soft drink manufacturer.

Measuring the spread of the data

Measures of dispersion identify how spread out a data set is, relative to the center. This provides a way of determining if the members of a data set tend to be very close to each other or if they tend to be widely scattered. Some of the most important measures of dispersion are

- Variance
- Standard deviation
- Percentiles
- Quartiles
- Interquartile range (IQR)

The *variance* is a measure of the average squared difference between the elements of a data set and the mean. The larger the variance, the more "spread out" the data is. Variance is often used as a measure of risk in business applications; for example, it can be used to show how much uncertainty there is over the returns on a stock.

The s*tandard deviation* is the square root of the variance, and is more commonly used than the variance (because the variance is expressed in squared units). For example, the variance of a series of gas prices is measured in squared dollars, which is difficult to interpret. The corresponding standard deviation is measured in dollars, which is much more intuitively clear.

Percentiles divide a data set into 100 equal parts, each consisting of 1 percent of the total. For example, if a student's score on a standardized exam is in the 80th percentile, then the student scored as well as or better than 80 percent of the other

students who took the exam. A *quartile* is a special type of percentile; it divides a data set into four equal parts, each consisting of 25 percent of the total. The first quartile is the 25th percentile of a data set, the second quartile is the 50th percentile, and the third quartile is the 75th percentile. The *interquartile range* identifies the middle 50 percent of the observations in a data set; it equals the difference between the third and the first quartiles.

Determining the relationship between two variables

For some applications, you need to understand the relationship between two variables. For example, if an investor wants to understand the risk of a portfolio of stocks, it's essential to properly measure how closely the returns on the stocks track each other. You can determine the relationship between two variables with two measures of *association:* covariance and correlation.

Covariance is used to measure the tendency for two variables to rise above their means or fall below their means at the same time. For example, suppose that a bioengineering company finds that increasing research and development expenditures typically leads to an increase in the development of new patents. In this case, R&D spending and new patents would have a positive covariance. If the same company finds that rising labor costs typically reduce corporate profits, then labor costs and profits would have a negative covariance. If the company finds that profits are not related to the average daily temperature, then these two variables will have a covariance that is very close to zero.

Correlation is a closely related measure. It's defined as a value between −1 and 1, so interpreting the correlation is easier than the covariance. For example, a correlation of 0.9 between two variables would indicate a very strong positive relationship, whereas a correlation of 0.2 would indicate a fairly weak but positive relationship. A correlation of −0.8 would indicate a very strong negative relationship; a correlation of −0.3 would indicate a weak negative relationship. A correlation of 0 would show that two variables have no linear relationship between them.

Probability: The Foundation of All Statistical Analysis

Probability theory provides a mathematical framework for measuring uncertainty. This area is important for business applications since much statistical analysis is strongly related to probability theory. Understanding probability theory provides fundamental insights into all the statistical methods used in this book.

Probability is heavily based on the notion of *sets.* A set is a collection of objects. These objects may be numbers, colors, flavors, and so on. This chapter focuses on sets of numbers that may represent prices, rates of return, and so forth. Several mathematical operations may be applied to sets — union, intersection, and complement, for example.

The union of two sets is a new set that contains all the elements in the original two sets. The intersection of two sets is a set that contains only the elements contained in *both* of the two original sets (if any). The complement of a set is a set containing elements that are *not* in the original set. For example, the complement of the set of black cards in a standard deck is the set containing all red cards.

Probability theory is based on a model of how random outcomes are generated, known as a *random experiment.* Outcomes are generated in such a way that all *possible* outcomes are known in advance, but the *actual* outcome isn't known. The following rules help you determine the probability of specific outcomes occurring:

- The addition rule
- The multiplication rule
- The complement rule

You use the addition rule to determine the probability of a union of two sets. The multiplication rule is used to determine the probability of an intersection of two sets. The complement rule is used to identify the probability that the outcome of a random experiment will *not* be an element in a specified set.

Random variables

A *random variable* assigns numerical values to the outcomes of a random experiment. For example, when you flip a coin twice, you're performing a random experiment because

- All possible outcomes are known in advance.
- The actual outcome isn't known in advance.

The experiment consists of two *trials*. On each trial, the outcome must be a "head" or a "tail."

Assume that a random variable X is defined as the number of "heads" that turn up during the course of this experiment. X assigns values to the outcomes of this experiment as follows:

Outcome	X
{TT}	0
{HT, TH}	1
{HH}	2

where:

T represents a tail on a single flip

H represents a head on a single flip

TT represents two consecutive tails

HT represents a head followed by a tail

TH represents a tail followed by a head

HH represents two consecutive heads

X assigns a value of 0 to the outcome TT because no heads turned up. *X* assigns a value of 1 to both HT and TH because one head turned up in each case. Similarly, *X* assigns a value of 2 to HH because two heads turned up.

Probability distributions

A *probability distribution* is a formula or a table used to assign probabilities to each possible value of a random variable *X.* A probability distribution may be *discrete*, which means that *X* can assume one of a finite (countable) number of values, or *continuous*, in which case *X* can assume one of an infinite (uncountable) number of different values.

For the coin-flipping experiment from the previous section, the probability distribution of *X* can be a simple table that shows the probability of each possible value of *X*, written as *P*(*X*):

X	P(X)
0	0.25
1	0.50
2	0.25

The probability that $X = 0$ (that no heads turn up) equals 0.25 because this experiment has four equally likely outcomes: HH, HT, TH, and TT and in only one of those cases will there be no heads. You compute the other probabilities in a similar manner.

Discrete probability distributions

Several specialized discrete probability distributions are useful for specific applications. For business applications, two frequently used discrete distributions are:

- Binomial
- Poisson

You use the *binomial distribution* to compute probabilities for a process where only one of two possible outcomes may occur on each trial. You can use the *Poisson distribution* to measure the probability that a given number of events will occur during a given time frame.

Continuous probability distributions

Many continuous distributions may be used for business applications; one of the most widely used is the normal distribution. The *normal distribution* is useful for a wide array of applications in many disciplines. In business applications, variables such as stock returns are often assumed to follow the normal distribution. The normal distribution is characterized by a *bell-shaped curve,* and areas under this curve represent probabilities. The bell-shaped curve is shown in Figure 1-5.

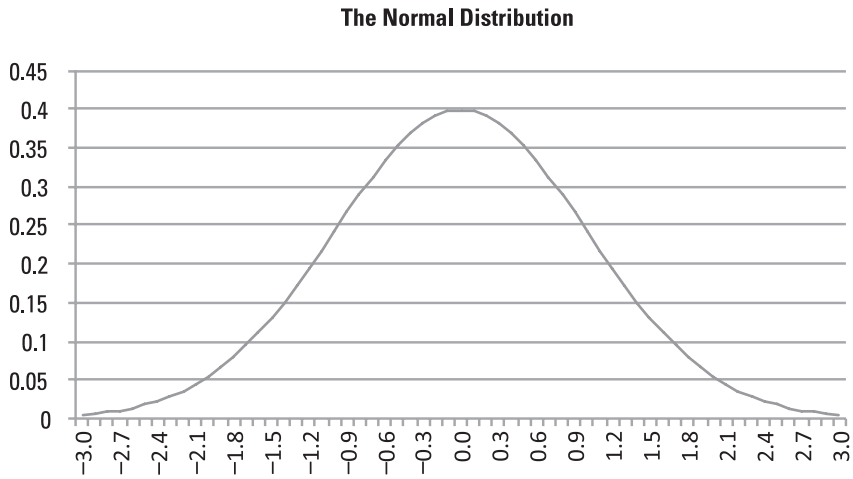

FIGURE 1-5: The bell-shaped curve of the normal distribution.

The normal distribution has many convenient statistical properties that make it a popular choice for statistical modeling. One of these properties is known as *symmetry*, the idea that the probabilities of values below the mean are matched by the probabilities of values that are equally far above the mean.

Using Sampling Techniques and Sampling Distributions

Sampling is used to determine the estimated properties of a population from sample data. A *population* is a collection of data that someone has an interest in studying. A sample is a selection of data randomly chosen from a population. For example, if a university is interested in analyzing the distribution of grade point averages (GPAs) among its MBA students, the population of interest would be the GPAs of every MBA student at the university; a sample would consist of the GPAs of a set of randomly chosen MBA students.

Several approaches can be used for choosing samples; a sample is a *subset* of the underlying population.

A *statistic* is a summary measure of a sample, while a *parameter* is a summary measure of a population. The properties of a statistic can be determined with a *sampling distribution* — a special type of probability distribution that describes the properties of a statistic.

The *central limit theorem* (CLT) gives the conditions under which the mean of a sample follows the normal distribution:

- The underlying population is normally distributed.
- The sample size is "large" (at least 30).

Statistical Inference: Drawing Conclusions from Data

Statistical inference refers to the process of drawing conclusions about a population from randomly chosen samples. In the following sections, I discuss two techniques used for statistical inference: confidence intervals and hypothesis testing.

Confidence intervals

A *confidence interval* is a set of values that's expected to contain the value of a population parameter with a specified level of confidence (such as 90 percent, 95 percent, 99 percent, and so on). For example, you can construct a confidence interval for the population mean by following these steps:

1. Estimate the value of the population mean by calculating the mean of a randomly chosen sample (known as the sample mean).
2. Calculate the lower limit of the confidence interval by subtracting a margin of error from the sample mean.
3. Calculate the upper limit of the confidence interval by adding the same margin of error to the sample mean.

The margin of error depends on the size of the sample used to construct the confidence interval, whether the population standard deviation is known, and the level of confidence chosen.

The resulting interval is known as a confidence interval. A confidence interval is constructed with a specified level of probability. For example, suppose you draw a sample of stocks from a portfolio, and you construct a 95 percent confidence interval for the mean return of the stocks in the entire portfolio:

(lower limit, upper limit) = (0.02, 0.08)

The returns on the entire portfolio are the population of interest. The mean return in each sample drawn is an *estimate* of the population mean. The sample mean will be slightly different each time a new sample is drawn, as will the confidence interval. If this process is repeated 100 times, 95 of the resulting confidence intervals will contain the true population mean.

Hypothesis testing

Hypothesis testing is a procedure for using sample data to draw conclusions about the characteristics of the underlying population.

The procedure begins with a statement, known as the *null hypothesis.* The null hypothesis is assumed to be true unless sufficient evidence against it is found. An *alternative hypothesis* — the result accepted if the null hypothesis is rejected — is also stated.

You calculate a *test statistic,* and you compare it with a *critical value* (or values) to determine whether the null hypothesis should be rejected. The specific test

statistic and critical value(s) depend on which population parameter is being tested, the size of the sample being used, and other factors.

If the test statistic is too extreme (for example, it's too large compared with the critical value[s]) the null hypothesis is rejected in favor of the alternative hypothesis; otherwise, the null hypothesis is *not* rejected.

TECHNICAL STUFF

If the null hypothesis isn't rejected, this doesn't necessarily mean that it's true; it simply means that there is not enough evidence to justify rejecting it.

Hypothesis testing is a general procedure and can be used to draw conclusions about many features of a population, such as its mean, variance, standard deviation, and so on.

Simple regression analysis

Regression analysis uses sample data to estimate the strength and direction of the relationship between two or more variables. *Simple regression analysis* estimates the relationship between a dependent variable (*Y*) and a single independent variable (*X*).

For example, suppose you're interested in analyzing the relationship between the annual returns of the Standard & Poor's (S&P) 500 Index and the annual returns of Apple stock. You can assume that the returns of Apple stock are related to the returns to the S&P 500 because the index is a reflection of the overall strength of the economy. The returns of Apple stock may be treated as the dependent variable (Y) and the returns of the S&P 500 may be treated as the independent variable (X). You can use regression analysis to measure the numerical relationship between the S&P 500 and Apple stock.

Simple regression analysis is based on the assumption that a linear relationship occurs between X and Y. A linear relationship takes this form:

$$Y = mX + b$$

Y is the dependent variable, *X* is the independent variable, *m* is the slope, and *b* is the intercept. The slope tells you how much *Y* changes due to a specific change in *X*; the intercept tells you what the value of *Y* would be if *X* had a value of zero.

The goal of regression analysis is to find a line that best fits or explains the data. The population regression line is written as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In this equation, Y_i is the dependent variable, X_i is the independent variable, β_0 is the intercept, β_1 is the slope, and ε_i is an error term.

A *sample* regression line, estimated from the data, is written as follows:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Here, $\hat{Y}_i$ is the estimated value of Y_i, $\hat{\beta}_0$ is the estimated value of β_0, $\hat{\beta}_1$ is the estimated value of β_1 and X_i is the independent variable.

The sample regression line shows the estimated relationship between Y and X; you can use this relationship to determine how Y responds to a given change in *X*. You can also use it to *forecast* future values of *Y* based on assumed values of *X*.

After estimating the sample regression line, the results are subjected to a series of tests to determine whether the equation is valid. If the equation isn't valid, you reject the results and try a new model.