3

Descriptive Statistics: Statistical Measurements and Probability Distributions

Figure 3.1 Statistics and data science play significant roles in stock market analysis, offering insights into market trends and risk assessment for better financial decision-making. (credit: modification of work “That was supposed to be going up, wasn't it?” by Rafael Matsunaga/Flickr, CC BY 2.0)

Chapter Outline

Introduction

Statistical analysis is the science of collecting, organizing, and interpreting data to make decisions. Statistical analysis lies at the core of data science, with applications ranging from consumer analysis (e.g., credit scores, retirement planning, and insurance) to government and business concerns (e.g., predicting inflation rates) to medical and engineering analysis.

Statistical analysis is an essential aspect of data science, involving the systematic collection, organization, and interpretation of data for decision-making. It serves as the foundation for various applications in consumer analysis such as credit scoring, retirement planning, and insurance as well as in government and business decision-making processes such as inflation rate prediction and marketing strategies. As a consumer, statistical analysis plays a significant role in various decision-making processes. For instance, when considering a large financial decision such as purchasing a house, the probability of interest rate fluctuations and their impact on mortgage financing must be taken into account.

Part of statistical analysis involves **descriptive statistics**, which refers to the collection, organization, summarization, and presentation of data using various graphs and displays. Once data is collected and summarized using descriptive statistics methods, the next step is to analyze the data using various probability tools and probability distributions in order to come to conclusions about the dataset and formulate predictions that will be useful for planning and estimation purposes.

Descriptive statistics includes measures of the center and dispersion of data, measures of position, and generation of various graphical displays. For example, a human resources administrator might be interested in generating some statistical measurements for the distribution of salaries at a certain company. This might involve calculating the mean salary, the median salary, and standard deviation, among others. The administrator might want to present the data to other employees, so generating various graphical displays such as histograms, box plots, and scatter plots might also be appropriate.

The human resources administrator might also want to derive estimates and predictions about the average salary in the company five years into the future, taking into account inflation effects, or they might want to create a model to predict an individual employee's salary based on the employee's years of experience in the field.

Such analyses are based on the statistical methods and techniques we will discuss in this chapter, building on the introduction to statistical analysis presented in What Are Data and Data Science? and Collecting and Preparing Data. Statistical analysis utilizes a variety of technological tools to automate statistical calculations and generate graphical displays. This chapter also will demonstrate the use of Python to generate results and graphs. The supplementary material at the end of this book will show the same using Excel (Appendix A) and R (Appendix B).

In this chapter, you will also study probability concepts and probability distributions. Probability theory helps us deal with quantifying uncertainty, which is always inherent in real-world data. We will see that in real-world datasets we often have to deal with noise and randomness that will be analyzed using statistical analyses.

Probability analysis provides the tools to model, understand, and quantify uncertainties, allowing data scientists to make informed decisions from data. Probability theory also forms the basis for different types of analyses, such as confidence intervals and hypothesis testing, which will be discussed further in Inferential Statistics and Regression Analysis. Such methods rely on probability models to make predictions and detect patterns. In addition, machine learning (discussed in Time Series and Forecasting) uses probabilistic models to help represent uncertainty and make predictions based on the collected data. Probability analysis helps data scientists to select data models, interpret the results, and assess the reliability of these conclusions. For a more detailed review of statistical concepts, please refer to Introductory Statistics 2e (https://openstax.org/books/introductory-statistics-2e/pages/1-introduction).

3.1 Measures of Center

Learning Outcomes

By the end of this section, you should be able to:

- 3.1.1 Define and calculate mean, trimmed mean, median, and mode for a dataset.
- 3.1.2 Determine the effect of outliers on the mean and the median.
- 3.1.3 Use Python to calculate measures of center for a dataset.

Measures of center are statistical measurements that provide a central, or typical, representation of a dataset. These measures can help indicate where the bulk of the data is concentrated and are often called the data's **central tendency**. The most widely used measures of the center of a dataset are the *mean* (average), the *median*, and the *mode*.

Mean and Trimmed Mean

The **mean**, or average, (sometimes referred to as the *arithmetic mean*) is the most commonly used measure of the center of a dataset. Sometimes the mean can be skewed by the presence of **outliers**, or data values that are significantly different as compared to the remainder of the dataset. In these instances, the *trimmed mean* is often used to provide a more representative measure of the center of the dataset, as we will discuss in the following section.

Mean

To calculate the mean, add the values of all the items in a dataset and divide by the number of items. For example, if the scores on your last three exams were 87, 92, and 73, then the mean score would be $\frac{87+92+73}{3} = 84$. If you had a large number of data values, you would proceed in the same way. For example, to calculate the mean value of 50 exam scores, add the 50 scores together and divide by 50. If the 50 scores add up to 4,050, for example, the mean score is $\frac{4050}{50}$, or 81.

In data science applications, you will encounter two types of datasets: sample data and population data. **Population data** represents all the outcomes or measurements that are of interest. **Sample data** represents outcomes or measurements collected from a subset, or part, of the population of interest. Of course in many applications, collecting data from an entire population is not practical or feasible, and so we often rely on sample data.

The notation $\bar{x}$ is used to indicate the **sample mean**, where the mean is calculated based on data taken from a sample. The notation $\sum x$ is used to denote the sum of the data values, and n is used to indicate the number of data values in the sample, also known as the **sample size**.

The sample mean can be calculated using the following formula:

$$\bar{x} = \frac{\sum x}{n}$$

The notation μ is used to indicate the **population mean**, where the mean is calculated based on data taken from the entire population, and N is used to indicate the number of data values in the population, also known as the **population size**. The population mean can be calculated using the following formula:

$$\mu = \frac{\sum x}{N}$$

The mean can also be determined by its frequency distribution. For every unique data value in the dataset, the **frequency distribution** gives the number of times, or **frequency**, that this unique value appears in the dataset. In this type of situation, the mean can be calculated by multiplying each distinct value by its frequency, summing these values, and then dividing this sum by the total number of data values. Here is the corresponding formula for the sample mean using the frequency distribution:

$$\bar{x} = \frac{\sum x \cdot f}{n}$$

When all the values in the dataset are unique, this reduces to the previous formula given for the sample mean.

EXAMPLE 3.1

Problem

During a clinical trial, a sample is taken of 10 patients and pulse rates are measured in beats per minute:

68, 92, 76, 51, 65, 83, 94, 72, 88, 59

Calculate the mean pulse rate for this sample.

Solution

Add the 10 data values, and the sum is 748. Divide this sum by the number of data values, which is 10. The result is:

$$\bar{x} = \frac{\sum x}{n} = \frac{748}{10} = 74.8$$

EXAMPLE 3.2

Problem

A college professor records the ages of 25 students in a data science class as shown:

Student Age	Number of Students (Frequency)
19	3
20	4
21	8
22	6
23	2
27	1
31	1
Total	**25**

Calculate the mean age for this sample of students.

Solution

Substitute the values from the table in the following formula:

$$\bar{x} = \frac{\sum x \cdot f}{n} = \frac{19 \cdot 3 + 20 \cdot 4 + 21 \cdot 8 + 22 \cdot 6 + 23 \cdot 2 + 27 \cdot 1 + 31 \cdot 1}{25} = \frac{541}{25} = 21.64$$

Trimmed Mean

A **trimmed mean** helps mitigate the effects of outliers, which are data values that are significantly different from most of the other data values in the dataset. In the dataset given in Example 3.1, a pulse rate of 35 or 120 would be considered outlier data values since these pulse rates are significantly different as compared to the rest of the values in Example 3.1. We will see that there is a formal method for determining outliers, and in fact there are several methods to identify outliers in a dataset.

The presence of outlier data values tends to disproportionally skew the mean and produce a potentially misleading result for the mean.

To calculate the trimmed mean for a dataset, first sort the data in ascending order (from smallest to largest). Then decide on a certain percentage of data values to be deleted from the lower and upper ends of the dataset. This might represent the extent of outliers in the dataset; trimmed mean percentages of 10% and 20% are common. Then delete the specified percentage of data values from both the lower end and upper end of the dataset. Then find the mean for the remaining undeleted data values.

As an example, to calculate a 10% trimmed mean, first sort the data values from smallest to largest. Then delete the lower 10% of the data values and delete the upper 10% of the data values. Then calculate the mean for the resulting dataset. Any outliers would tend to be deleted as part of the trimmed mean calculation, and thus the trimmed mean would then be a more representative measure of the center of the data for datasets containing outliers.

EXAMPLE 3.3

Problem

A real estate agent collects data on a sample of recently sold homes in a certain neighborhood, and the data are shown in the following dataset:

397900, 452600, 507400, 488300, 623400, 573200, 1689300, 403890, 612300, 599000, 2345800, 499000, 525000, 675000, 385000

1. Calculate the mean of the dataset.
2. Calculate a 20% trimmed mean rate for the dataset.

Solution

1. For the mean, add the 15 data values, and the sum is 10,777,090. Divide this sum by the number of data values, which is 15. The result is:

$$\bar{x} = \frac{\sum x}{n} = \frac{10{,}777{,}090}{15} = 718{,}472.70$$

2. For the trimmed mean, first order the data from smallest to largest. The sorted dataset is:

   385000, 397900, 403890, 452600, 488300, 499000, 507400, 525000, 573200, 599000, 612300, 623400, 675000, 1689300, 2345800

   Twenty percent of 15 data values is 3, and this indicates that 3 data values are to be deleted from each of the lower end and upper end of the dataset. The resulting 9 undeleted data values are:

   452600, 488300, 499000, 507400, 525000, 573200, 599000, 612300, 623400

   Then find the mean for the remaining data values. The sum of these 9 data values is 4,880,200. Divide this sum by the number of data values (9). The result is:

$$\bar{x} = \frac{\sum x}{n} = \frac{4{,}880{,}200}{9} = 542{,}244.40$$

Notice how the mean calculated in Part (a) is significantly larger as compared to the trimmed mean calculated in Part (b). The reason is the presence of several large outlier home prices. Once these outlier data values are removed by the trimmed mean calculation, the resulting trimmed mean is more representative of the typical home price in this neighborhood as compared to the mean.

Median

The median provides another measure of central tendency for a dataset. The **median** is generally a better measure of the central tendency when there are outliers (extreme values) in the dataset. Since the median focuses on the middle value of the ordered dataset, the median is preferred when outliers are present because the median is not affected by the numerical values of the outliers.

To determine the median of a dataset, first order the data from smallest to largest, and then find the middle value in the ordered dataset. For example, to find the median value of 50 exam scores, find the score that splits the data into two equal parts. The exam scores for 25 students will be below the median, and 25 students will have exam scores above the median.

If there is an odd number of data values in the dataset, then there will be one data value that represents the middle value, and this is the median. If there is an even number of data values in the dataset, then to find the median, add the two middle values together and divide by 2 (this is essentially finding the mean of the two middle values in the dataset).

EXAMPLE 3.4

Problem

The same dataset of pulse rates from Example 3.1 is:

68, 92, 76, 51, 65, 83, 94, 72, 88, 59

Calculate the median pulse rate for this sample.

Solution

First, order the 10 data values from smallest to largest. Divide this sum by the number of data values, which is 10. The result is:

51, 59, 65, 68, 72, 76, 83, 88, 92, 94

Since there is an even number of data values, add the two middle values together and divide by 2.
The two middle values are 72 and 76.

$$\text{Median} = \frac{72 + 76}{2} = \frac{148}{2} = 74$$

You can also quickly find the sample median of a dataset as follows.

Let n represent the number of data values in the sample.

- If n is odd, then the median is the data value in position $\frac{n+1}{2}$.
- If n is even, the median is the mean of the observations in position $\frac{n}{2}$ and position $\frac{n}{2} + 1$.

For example, let's say a dataset has 25 data values. Since n is odd, to identify the position of the median, calculate $\frac{n+1}{2}$, which is $\frac{25+1}{2}$, or 13. This indicates that the median is located in the 13th data position.

As another example, let's say a dataset has 100 data values. Since n is even, to identify the position of the median, calculate $\frac{n}{2}$, which is $\frac{100}{2}$, which is 50, and also calculate $\frac{n}{2} + 1$, which is $50 + 1$, which is 51. This indicates that the median is calculated as the mean of the 50th and 51st data values.

Mode

Another measure of center is the **mode**. The mode is the data value that occurs with the greatest frequency. If there are no repeating data values in a dataset, then there is no mode. If two data values occur with same greatest frequency, then there are two modes, and we say the data is bimodal. For example, assume that the weekly closing stock price for a technology stock, in dollars, is recorded for 20 consecutive weeks as follows:

50, 53, 59, 59, 63, 63, 72, 72, 72, 72, 72, 76, 78, 81, 83, 84, 84, 84, 90, 93

To find the mode, determine the most frequent score, which is 72, which occurs five times. Thus, the mode of this dataset is 72.

The mode can also be applied to non-numeric (qualitative) data, whereas the mean and the median can only be applied for numeric (quantitative) data. For example, a restaurant manager might want to determine the mode for responses to customer surveys on the quality of the service of a restaurant, as shown in Table 3.1.

Customer Service Rating	Number of Respondents
Excellent	267
Very Good	410
Good	392
Fair	107
Poor	18

Table 3.1 Customer Survey Results for Customer Survey Rating

Based on the survey responses, the mode is the Customer Service Rating of "Very Good," since this is the data value with the greatest frequency.

Influence of Outliers on Measures of Center

As mentioned earlier, when outliers are present in a dataset, the mean may not represent the center of the dataset, and the median will provide a better measure of center. The reason is that the median focuses on the middle value of the ordered dataset. Thus, any outliers at the lower end of the dataset or any outliers at the upper end of the dataset will not affect the median. Note: A formal method for identifying outliers is presented in Measures of Position when measures of position are discussed. The following example illustrates the point that the median is a better measure of central tendency when potential outliers are present.

EXAMPLE 3.5

Problem

Suppose that in a small company of 40 employees, one person earns a salary of $3 million per year, and the other 39 individuals each earn $40,000. Which is the better measure of center: the mean or the median?

Solution

The mean, in dollars, would be arrived at mathematically as follows:

$$\bar{x} = \frac{3{,}000{,}000 + 39(40{,}000)}{40} = 114{,}000$$

However, the median would be $40,000 since this is the middle data value in the ordered dataset. There are 39 people who earn $40,000 and one person who earns $3,000,000.

Notice that the mean is not representative of the typical value in the dataset since $114,000 is not reflective of the average salary for most employees (who are earning $40,000). The median is a much better measure of the "average" than the mean in this case because 39 of the values are $40,000 and one is $3,000,000. The data value of $3,000,000 is an outlier. The median result of $40,000 gives us a better sense of the center of the dataset.

Using Python for Measures of Center

We learned in What Are Data and Data Science? how the `DataFrame.describe()` method is used to

summarize data. Recall that the method `describe()` is defined for a DataFrame type object so should be called upon a DataFrame type variable (e.g. given a DataFrame `d`, use `d.describe()`).

Figure 3.2 shows the output of `DataFrame.describe()` on the "Movie Profit" dataset we used in What Are Data and Data Science?, movie_profit.csv. The mean and 50% quartile show the average and median of each column. For example, the average worldwide gross earnings are $410.14 million, and the median earnings are $309.35 million. Note that average and/or median of some columns are not as meaningful as the others. The first column—Unnamed: 0—was simply used as an identifier of each item in the dataset, so the average and mean of this column is not quite useful. `DataFrame.describe()` still computes the values because it can (and it does *not* care which column is meaningful to do so or not).

	Unnamed: 0	Rating	Duration	US_Gross_Million	Worldwide_Gross_Million
count	966.00000	966.000000	966.000000	966.000000	966.000000
mean	483.50000	6.814286	117.506211	156.158975	410.140600
std	279.00448	0.894383	21.615612	110.629617	294.758791
min	1.00000	3.300000	69.000000	0.010000	176.600000
25%	242.25000	6.200000	101.250000	90.832500	223.277500
50%	483.50000	6.800000	116.000000	129.245000	309.345000
75%	724.75000	7.400000	130.000000	187.090000	472.645000
max	966.00000	9.200000	238.000000	936.660000	2847.400000

Figure 3.2 The Output of `DataFrame.describe()` with the Movie Profit Dataset

EXPLORING FURTHER

Working with Python

See the Python website (https://openstax.org/r/python) for more details on using, installing, and working with Python. See this additional documentation, for more specific information on the statistics module (https://openstax.org/r/docspython).

3.2 Measures of Variation

Learning Outcomes

By the end of this section, you should be able to:

- 3.2.1 Define and calculate the range, the variance, and the standard deviation for a dataset.
- 3.2.2 Use Python to calculate measures of variation for a dataset.

Providing some measure of the spread, or *variation*, in a dataset is crucial to a comprehensive summary of the dataset. Two datasets may have the same mean but can exhibit very different spread, and so a measure of dispersion for a dataset is very important. While measures of central tendency (like mean, median, and mode) describe the center or average value of a distribution, measures of dispersion give insights into how much individual data points deviate from this central value.

The following two datasets are the exam scores for a group of three students in a biology course and in a statistics course.

Dataset A: Exam scores for students in a biology course: 40, 70, 100
Dataset B: Exam scores for students in a statistics course: 69, 70, 71

Notice that the mean score for both Dataset A and Dataset B is 70.

However, the datasets are significantly different from one another:

Dataset A has larger variability where one student scored 30 points below the mean and another student scored 30 points above the mean.
Dataset B has smaller variability where the exam scores are much more tightly clustered around the mean of 70.

This example illustrates that publishing the mean of a dataset is often inadequate to fully communicate the characteristics of the dataset. Instead, data scientists will typically include a measure of variation as well.

The three primary measures of variability are range, variance, and standard deviation, and these are described next.

Range

Range is a measure of dispersion for a dataset that is calculated by subtracting the minimum from the maximum of the dataset:

$$\text{Range} = \text{Max} - \text{Min}$$

Range is a straightforward calculation but makes use of only two of the data values in a dataset. The range can also be affected by outliers.

EXAMPLE 3.6

Problem

Calculate the range for Dataset A and Dataset B:

Dataset A: Exam scores for students in a biology course: 40, 70, 100
Dataset B: Exam scores for students in a statistics course: 69, 70, 71

Solution

For Dataset A, the maximum data value is 100 and the minimum data value is 40.
The range is then calculated as:

$$\begin{aligned} \text{Range} &= \text{Max} - \text{Min} \\ \text{Range} &= 100 - 40 \\ \text{Range} &= 60 \end{aligned}$$

For Dataset B, the maximum data value is 71 and the minimum data value is 69.
The range is then calculated as:

$$\begin{aligned} \text{Range} &= \text{Max} - \text{Min} \\ \text{Range} &= 71 - 69 \\ \text{Range} &= 2 \end{aligned}$$

The range clearly indicates that there is much less spread in Dataset B as compared to Dataset A.

One drawback to the use of the range is that it doesn't take into account every data value. The range only uses two data values from the dataset: the minimum (min) and the maximum (max). Also the range is influenced by outliers since an outlier might appear as a minimum or maximum data value and thus skew the results. For these reasons, we typically use other measures of variation, such as variance or standard deviation.

Variance

The **variance** provides a measure of the spread of data values by using the squared deviations from the mean. The more the individual data values differ from the mean, the larger the variance.

A financial advisor might use variance to determine the volatility of an investment and therefore help guide financial decisions. For example, a more cautious investor might opt for investments with low volatility.

The formula used to calculate variance also depends on whether the data is collected from a sample or a population. The notation s^2 is used to represent the *sample variance*, and the notation σ^2 is used to represent the *population variance*.

Formula for the sample variance:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Formula for the population variance:

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$$

In these formulas:
x represents the individual data values
$\bar{x}$ represents the sample mean
n represents the sample size
μ represents the population mean
N represents the population size

ALTERNATE FORMULA FOR VARIANCE

An alternate formula for the variance is available. It is sometimes used for more efficient computations:

$$\sigma^2 = \frac{\sum x^2}{N} - \mu^2$$

In the formulas for sample variance and population variance, notice the denominator for the sample variance is $n - 1$, whereas the denominator for the population variance is N. The use of $n - 1$ in the denominator of the sample variance is used to provide the best estimate for the population variance, in the sense that if repeated samples of size n are taken and the sample mean computed each time, then the average of those sample means will tend to the population mean as the number of repeated samples increase.

It is important to note that in many data science applications, population data is unavailable, and so we typically calculate the sample variance. For example, if a researcher wanted to estimate the percentage of smokers for all adults in the United States, it would be impractical to collect data from every adult in the United States.

Notice that the sample variance is a sum of squares. Its units of measurement are squares of the units of measurement of the original data. Since these square units are different than the units in the original data, this can be confusing. By contrast, standard deviation is measured in the same units as the original dataset, and thus the standard deviation is more commonly used to measure the spread of a dataset.

Standard Deviation

The **standard deviation** of a dataset provides a numerical measure of the overall amount of variation *in a dataset in the same units as the data*; it can be used to determine whether a particular data value is close to or

far from the mean, relative to the typical distance from the mean.

The standard deviation is always positive or zero. It is small when the data values are all concentrated close to the mean, exhibiting little variation, or spread. It is larger when the data values are spread out more from the mean, exhibiting more variation. A smaller standard deviation implies less variability in a dataset, and a larger standard deviation implies more variability in a dataset.

Suppose that we are studying the variability of two companies (A and B) with respect to employee salaries. The average salary for both companies is $60,000. For Company A, the standard deviation of salaries is $8,000, whereas the standard deviation for salaries for Company B is $19,000. Because Company B has a higher standard deviation, we know that there is more variation in the employee salaries for Company B as compared to Company A.

There are two different formulas for calculating standard deviation. Which formula to use depends on whether the data represents a sample or a population. The notation s is used to represent the sample standard deviation, and the notation σ is used to represent the population standard deviation. In the formulas shown, $\bar{x}$ is the sample mean, μ is the population mean, n is the sample size, and N is the population size.

Formula for the sample standard deviation:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Formula for the population standard deviation:

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$$

Notice that the sample standard deviation is calculated as the square root of the variance. This means that once the sample variance has been calculated, the sample standard deviation can then be easily calculated as the square root of the sample variance, as in Example 3.7.

EXAMPLE 3.7

Problem

A biologist calculates that the sample variance for the amount of plant growth for a sample of plants is 8.7 cm^2. Calculate the sample standard deviation.

Solution

The sample standard deviation (s) is calculated as the square root of the variance.

$$s = \sqrt{s^2} = \sqrt{8.7} = 2.9 \text{ cm}$$

EXAMPLE 3.8

Problem

Assume the sample variance (s^2) for a dataset is calculated as 42.2. Based on this, calculate the sample standard deviation.

Solution

The sample standard deviation (s) is calculated as the square root of the variance.

$$s = \sqrt{s^2} = \sqrt{42.2} = 6.5 \text{ years}$$

This result indicates that the standard deviation is about 6.5 years.

Notice that the sample variance is the square of the sample standard deviation, so if the sample standard deviation is known, the sample variance can easily be calculated.

USE OF TECHNOLOGY FOR CALCULATING MEASURES OF VARIABILITY

Due to the complexity of calculating variance and standard deviation, technology is typically utilized to calculate these measures of variability. For example, refer to the examples shown in Coefficient of Variation on using Python for measures of variation.

Coefficient of Variation

A data scientist might be interested in comparing variation with different units of measurement of different means, and in these scenarios the **coefficient of variation (CV)** can be used. The coefficient of variation measures the variation of a dataset by calculating the standard deviation as a percentage of the mean. Note: coefficient of variation is typically expressed in a percentage format.

$$\begin{aligned} \text{CV} &= \frac{\sigma}{\mu} \times 100\% \\ \text{Sample CV} &= \frac{s}{\bar{x}} \times 100\% \end{aligned}$$

EXAMPLE 3.9

Problem

Compare the relative variability for Company A versus Company B using the coefficient of variation, based on the following sample data:

Company A: $\text{Sample Mean} = \$68{,}000,\ \text{Sample Standard Deviation} = \$9{,}200$

Company B: $\text{Sample Mean} = \$71{,}000,\ \text{Sample Standard Deviation} = \$6{,}400$

Solution

Calculate the coefficient of variation for each company:

$$\text{CV for Company A} = \frac{s}{\bar{x}} \times 100\% = \frac{9{,}200}{68{,}000} \times 100\% = 13.5\%$$

$$\text{CV for Company B} = \frac{s}{\bar{x}} \times 100\% = \frac{6{,}400}{71{,}000} \times 100\% = 9.0\%$$

Company A exhibits more variability relative to the mean as compared to Company B.

Using Python for Measures of Variation

`DataFrame.describe()` computes standard deviation as well on each column of a dataset. The **std** lists the standard deviation of each column (See Figure 3.3).

	Unnamed: 0	Rating	Duration	US_Gross_Million	Worldwide_Gross_Million
count	966.00000	966.000000	966.000000	966.000000	966.000000
mean	483.50000	6.814286	117.506211	156.158975	410.140600
std	279.00448	0.894383	21.615612	110.629617	294.758791
min	1.00000	3.300000	69.000000	0.010000	176.600000
25%	242.25000	6.200000	101.250000	90.832500	223.277500
50%	483.50000	6.800000	116.000000	129.245000	309.345000
75%	724.75000	7.400000	130.000000	187.090000	472.645000
max	966.00000	9.200000	238.000000	936.660000	2847.400000

Figure 3.3 The Output of `DataFrame.describe()` with the Movie Profit Dataset

3.3 Measures of Position

Learning Outcomes

By the end of this section, you should be able to:

- 3.3.1 Define and calculate percentiles, quartiles, and z-scores for a dataset.
- 3.3.2 Use Python to calculate measures of position for a dataset.

Common measures of position include percentiles and quartiles as well as z-scores, all of which are used to indicate the relative location of a particular datapoint.

Percentiles

If a student scores 47 on a biology exam, it is difficult to know if the student did well or poorly compared to the population of all other students taking the exam. **Percentiles** provide a way to assess and compare the distribution of values and the position of a specific data point in relation to the entire dataset by indicating the percentage of data points that fall below it. Specifically, a percentile is a value on a scale of one hundred that indicates the percentage of a distribution that is equal to or below it. Let's say the student learns they scored in the 90th percentile on the biology exam. This percentile indicates that the student has an exam score higher than 90% of all other students taking the test. This is the same as saying that the student's score places the student in the top 10% of all students taking the biology test. Thus, this student scoring in the 90th percentile did very well on the exam, even if the actual score was 47.

To calculate percentiles, the data must be ordered from smallest to largest and then the ordered data divided into hundredths. If you score in the 80th percentile on an aptitude test, that does not necessarily mean that you scored 80% on the test. It means that 80% of the test scores are the same as or less than your score and the remaining 20% of the scores are the same as or greater than your score.

Percentiles are useful for comparing many types of values. For example, a stock market mutual fund might report that the performance for the fund over the past year was in the 80th percentile of all mutual funds in the peer group. This indicates that the fund performed better than 80% of all other funds in the peer group. This also indicates that 20% of the funds performed better than this particular fund.

To calculate percentiles for a *specific data value* in a dataset, first order the dataset from smallest to largest and count the number of data values in the dataset. Locate the measurement of interest and count how many data values fall below the measurement. Then the percentile for the measurement is calculated as follows:

$$\text{Percentile} = \frac{\text{number of data values below the measurement}}{\text{total number of data values}} \times 100\% = \frac{n}{N} \times 100\%$$

EXAMPLE 3.10

Problem

The following ordered dataset represents the scores of 15 employees on an aptitude test:

51, 63, 65, 68, 71, 75, 75, 77, 79, 82, 88, 89, 89, 92, 95

Determine the percentile for the employee who scored 88 on the aptitude test.

Solution

There are 15 data values in total, and there are 10 data values below 88.

$$\text{Percentile} = \frac{\text{number of data values below the measurement}}{\text{total number of data values}} \times 100\% = \frac{10}{15} \times 100\% = 67\text{th percentile}$$

Quartiles

While percentiles separate data into 100 equal parts, **quartiles** separate data into quarters, or four equal parts. To find the quartiles, first find the median, or second quartile. The first quartile, Q_1, is the middle value, or median, of the lower half of the data, and the third quartile, Q_3, is the middle value of the upper half of the data.

Note the following correspondence between quartiles and percentiles:

- The first quartile corresponds to the 25th percentile.
- The second quartile (which is the median) corresponds to the 50th percentile.
- The third quartile corresponds to the 75th percentile.

EXAMPLE 3.11

Problem

Consider the following ordered dataset, which represents the time in seconds for an athlete to complete a 40-yard run:

5.4, 6.0, 6.3, 6.8, 7.1, 7.2, 7.4, 7.5, 7.9, 8.2, 8.7

Solution

The median, or second quartile, is the middle value in this dataset, which is 7.2. Notice that 50% of the data values are below the median, and 50% of the data values are above the median. The lower half of the data values are 5.4, 6.0, 6.3, 6.8, 7.1. Note that these are the data values below the median. The upper half of the data values are 7.4, 7.5, 7.9, 8.2, 8.7, which are the data values above the median.

To find the first quartile, Q_1, locate the middle value of the lower half of the data (5.4, 6.0, 6.3, 6.8, 7.1). The middle value of the lower half of the dataset is 6.3. Notice that one-fourth, or 25%, of the data values are below this first quartile, and 75% of the data values are above this first quartile.

To find the third quartile, Q_3, locate the middle value of the upper half of the data (7.4, 7.5, 7.9, 8.2, 8.7). The middle value of the upper half of the dataset is 7.9. Notice that one-fourth, or 25%, of the data values are above this third quartile, and 75% of the data values are below this third quartile.

Thus, the quartiles Q_1, Q_2, Q_3 for this dataset are 6.3, 7.2, 7.9, respectively.

The **interquartile range (IQR)** is a number that indicates the spread of the middle half, or the middle 50%, of the data. It is the difference between the third quartile, Q_3, and the first quartile, Q_1.

$$\text{IQR} = Q_3 - Q_1$$

Note that the IQR provides a measure of variability that excludes outliers.

In Example 3.11, the IQR can be calculated as:

$$\text{IQR} = Q_3 - Q_1 = 7.9 - 6.3 = 1.6$$

Quartiles and the IQR can be used to flag possible outliers in a dataset. For example, if most employees at a company earn about $50,000 and the CEO of the company earns $2.5 million, then we consider the CEO's salary to be an outlier data value because this salary is significantly different from all the other salaries in the dataset. An outlier data value can also be a value much lower than the other data values in a dataset, so if one employee only makes $15,000, then this employee's low salary might also be considered an outlier.

To detect outliers, you can use the quartiles and the IQR to calculate a lower and an upper bound for outliers. Then any data values below the lower bound or above the upper bound will be flagged as outliers. These data values should be further investigated to determine the nature of the outlier condition and whether the data values are valid or not.

To calculate the lower and upper bounds for outliers, use the following formulas:

$$\begin{aligned} \text{Lower Bound for Outliers} &= Q_1 - (1.5 \cdot \text{IQR}) \\ \text{Upper Bound for Outliers} &= Q_3 + (1.5 \cdot \text{IQR}) \end{aligned}$$

These formulas typically use 1.5 as a cutoff value to identify outliers in a dataset.

EXAMPLE 3.12

Problem

Calculate the IQR for the following 13 home prices and determine if any of the home prices values are potential outliers. Data values are in US dollars.

389950, 230500, 158000, 479000, 639000, 114950, 5500000, 387000, 659000, 529000, 575000, 488800, 1095000

Solution

Order the data from smallest to largest.

114950, 158000, 230500, 387000, 389950, 479000, 488800, 529000, 575000, 639000, 659000, 1095000, 5500000

First, determine the median of the dataset. There are 13 data values, so the median is the middle data value, which is 488,800.

Next, calculate the Q_1 and Q_3.

For the first quartile, look at the data values below the median. The two middle data values in this lower half of the data are 230,500 and 387,000. To determine the first quartile, find the mean of these two data values.

For the third quartile, look at the data values above the median. The two middle data values in this upper half of the data are 639,000 and 659,000. To determine the third quartile, find the mean of these two data values.

$$Q_1 = \frac{230{,}500+387{,}000}{2} = 308{,}750$$
$$Q_3 = \frac{639{,}000+659{,}000}{2} = 649{,}000$$

Now, calculate the interquartile range (IQR):

$$\text{IQR} = 649{,}000 - 308{,}750 = 340{,}250$$

Calculate the value of 1.5 interquartile range (IQR):

$$(1.5)(\text{IQR}) = (1.5)(340{,}250) = 510{,}375$$

Calculate the lower and upper bound for outliers:

$$\text{Lower Bound} = Q_1 - (1.5)(\text{IQR}) = 308{,}750 - 510{,}375 = -201{,}625$$
$$\text{Upper Bound} = Q_3 + (1.5)(\text{IQR}) = 649{,}000 + 510{,}375 = 1{,}159{,}375$$

The lower bound for outliers is −201,625. Of course, no home price is less than −201,625, so no outliers are present for the lower end of the dataset.

The upper bound for outliers is 1,159,375. The data value of 5,500,000 is greater than the upper bound of 1,159,375. Therefore, the home price of $5,500,000 is a potential outlier. This is important because the presence of outliers could potentially indicate data errors or some other anomalies in the dataset that should be investigated. For example, there may have been a data entry error and a home price of $550,000 was erroneously entered as $5,500,000.

z-scores

The **z-score** is a measure of the position of an entry in a dataset that makes use of the mean and standard deviation of the data. It represents the number of standard deviations by which a data value differs from the mean. For example, suppose that in a certain neighborhood, the mean selling price of a home is $350,000 and the standard deviation is $40,000. A particular home sells for $270,000. Based on the selling price of this home, we can calculate the relative standing of this home compared to other home sales in the same neighborhood.

The corresponding z-score of a measurement considers the given measurement in relation to the mean and standard deviation for the entire population. The formula for a z-score calculation is as follows:

$$z = \frac{x - \mu}{\sigma}$$

Where:
x is the measurement
μ is the mean
σ is the standard deviation

Notice that when a measurement is below the mean, the corresponding z-score will be a negative value. If the measurement is exactly equal to the mean, the corresponding z-score will be zero. If the measurement is above the mean, the corresponding z-score will be a positive value.

z-scores can also be used to identify outliers. Since z-scores measure the number of standard deviations from the mean for a data value, a z-score of 3 would indicate a data value that is 3 standard deviations above the mean. This would represent a data value that is significantly displaced from the mean, and typically, a z-score less than −3 or a z-score greater than +3 can be used to flag outliers.

EXAMPLE 3.13

Problem

For the home example in Example 3.12, the x value is the home price of $270,000, the mean μ is $350,000, and the standard deviation σ is $40,000. Calculate the z-score.

Solution

The z-score can be calculated as follows:

$$z = \frac{x - \mu}{\sigma} = \frac{270{,}000 - 350{,}000}{40{,}000} = \frac{-80{,}000}{40{,}000} = -2$$

This z-score of −2 indicates that the selling price for this home is 2 standard deviations below the mean, which represents a data value that is significantly below the mean.

Using Python to Calculate Measures of Position for a Dataset

`DataFrame.describe()` computes different measures of position as well on each column of a dataset. See min, 25%, 50%, 75%, and max in Figure 3.4.

	Unnamed: 0	Rating	Duration	US_Gross_Million	Worldwide_Gross_Million
count	966.00000	966.000000	966.000000	966.000000	966.000000
mean	483.50000	6.814286	117.506211	156.158975	410.140600
std	279.00448	0.894383	21.615612	110.629617	294.758791
min	1.00000	3.300000	69.000000	0.010000	176.600000
25%	242.25000	6.200000	101.250000	90.832500	223.277500
50%	483.50000	6.800000	116.000000	129.245000	309.345000
75%	724.75000	7.400000	130.000000	187.090000	472.645000
max	966.00000	9.200000	238.000000	936.660000	2847.400000

Figure 3.4 The Output of `DataFrame.describe()` with the Movie Profit Dataset

3.4 Probability Theory

Learning Outcomes

By the end of this section, you should be able to:

- 3.4.1 Describe the basic concepts of probability and apply these concepts to real-world applications in data science.
- 3.4.2 Apply conditional probability and Bayes' Theorem.

Probability is a numerical measure that assesses the likelihood of occurrence of an event. Probability applications are ubiquitous in data science since many decisions in business, science, and engineering are based on probability considerations. We all use probability calculations every day as we decide, for instance, whether to take an umbrella to work, the optimal route for a morning commute, or the choice of a college major.

Basic Concepts of Probability

We have all used probability in one way or another on a day-to-day basis. Before leaving the house, you might

want to know the probability of rain. The probability of obtaining heads on one flip of a coin is one-half, or 0.5.

A data scientist is in interested in expressing probability as a number between 0 and 1 (inclusive), where 0 indicates impossibility (the event will not occur) and 1 indicates certainty (the event will occur). The probability of an event falling between 0 and 1 reflects the degree of uncertainty associated with the event.

Here is some terminology we will be using in probability-related analysis:

- An **outcome** is the result of a single trial in a probability experiment.
- The **sample space** is the set of all possible outcomes in a probability experiment.
- An **event** is some subset of the sample space. For example, an event could be rolling an even number on a six-sided die. This event corresponds to three outcomes, namely rolling a 2, 4, or 6 on the die.

To calculate probabilities, we can use several approaches, including relative frequency probability, which is based on actual data, and theoretical probability, which is based on theoretical conditions.

Relative Frequency Probability

Relative frequency probability is a method of determining the likelihood of an event occurring based on the observed frequency of its occurrence in a given sample or population. A data scientist conducts or observes a procedure and determines the number of times a certain Event A occurs. The probability of Event A, denoted as $P(A)$, is then calculated based on data that has been collected from the experiment, as follows:

$$P(A) = \frac{\text{number of times Event } A \text{ has occured}}{\text{number of times the procedure was repeated}}$$

EXAMPLE 3.14

Problem

A polling organization asks a sample of 400 people if they are in favor of increased funding for local schools; 312 of the respondents indicate they are in favor of increased funding. Calculate the probability that a randomly selected person will be in favor of increased funding for local schools.

Solution

Using the data collected from this polling, a total of 400 people were asked the question, and 312 people were in favor of increased school funding. The probability for a randomly selected person being in favor of increased funding can then be calculated as follows (notice that Event A in this example corresponds to the event that a person is in favor of the increased funding):

$$\begin{aligned} P(A) &= \text{Probability of In Favor} \\ &= \frac{\text{number of times Event } A \text{ has occurred}}{\text{number of times the procedure was repeated}} = \frac{312}{400} = 0.78 = 78\% \end{aligned}$$

EXAMPLE 3.15

Problem

A medical patient is told they need knee surgery, and they ask the doctor for an estimate of the probability of success for the surgical procedure. The doctor reviews data from the past two years and determines there were 200 such knee surgeries performed and 188 of them were successful. Based on this past data, the doctor calculates the probability of success for the knee surgery (notice that Event A in this example corresponds to the event that a patient has a successful knee surgery result).

Solution

Using the data collected from the past two years, there were 200 surgeries performed, with 188 successes. The probability can then be calculated as:

$$P(A) = \text{Probability of Success} = \frac{\text{number of times Event } A \text{ has occured}}{\text{number of times the procedure was repeated}} = \frac{188}{200} = 0.94$$

The doctor informs the patient that there is a 94% chance of success for the pending knee surgery.

Theoretical Probability

Theoretical probability is the method used when the outcomes in a probability experiment are equally likely—that is, under theoretical conditions.

The formula used for theoretical probability is similar to the formula used for **empirical probability**. Theoretical probability considers all the possible outcomes for an experiment that are known ahead of time so that past data is not needed in the calculation for theoretical probability.

$$\text{Theoretical Probability} = \frac{\text{number of outcomes for the event of interest}}{\text{total number of outcomes in the sample space}}$$

For example, the theoretical probability of rolling an even number when rolling a six-sided die is $\frac{3}{6}$ (which is $\frac{1}{2}$, or 0.5). There are 3 outcomes corresponding to rolling an even number, and there are 6 outcomes total in the sample space. Notice this calculation can be done without conducting any experiments since the outcomes are equally likely.

EXAMPLE 3.16

Problem

A student is working on a multiple-choice question that has 5 possible answers. The student does not have any idea about the correct answer, so the student randomly guesses. What is the probability that the student selects the correct answer?

Solution

Since the student is guessing, each answer choice is equally likely to be selected. There is 1 correct answer out of 5 possible choices. The probability of selecting the correct answer can be calculated as:

$$\begin{aligned}\text{Probability of Correct Answer} &= \frac{\text{number of outcomes for the event of interest}}{\text{total number of outcomes in the sample space}} \\ &= \frac{1}{5} = 0.20 = 20\%\end{aligned}$$

Notice in Example 3.16 that probabilities can be written as fractions, decimals, or percentages.

Also note that any probability must be between 0 and 1 inclusive. An event with a probability of zero will never occur, and an event with a probability of 1 is certain to occur. A probability greater than 1 is not possible, and a negative probability is not possible.

Complement of an Event

The **complement of an event** is the set of all outcomes in the sample space that are *not* included in the event. The complement of Event A is usually denoted by A' (A prime). To find the probability of the complement of Event A, subtract the probability of Event A from 1.

$$P(A') = 1 - P(A)$$

EXAMPLE 3.17

Problem

A company estimates that the probability that an employee will provide confidential information to a hacker is 0.1%. Determine the probability that an employee will not provide any confidential information during a hacking attempt.

Solution

Let Event A be the event that the employee will provide confidential information to a hacker. Then the complement of this Event A' is the event that an employee will not provide any confidential information during a hacking attempt.

$$\begin{aligned} P(A') &= 1 - P(A) \\ P(A') &= 1 - 0.001 = 0.999 \end{aligned}$$

There is a 99.9% probability that an employee will not provide any confidential information during a hacking attempt.

Conditional Probability and Bayes' Theorem

Data scientists are often interested in determining conditional probabilities, or the occurrence of one event that is conditional or dependent on another event. For example, a medical researcher might be interested to know if an asthma diagnosis for a patient is dependent on the patient's exposure to air pollutants. In addition, when calculating conditional probabilities, we can sometimes revise a probability estimate based on additional information that is obtained. As we'll see in the following section, Bayes' Theorem allows new information to be used to refine a probability estimate.

Conditional Probability

A **conditional probability** is the probability of an event given that another event has already occurred. The notation for conditional probability is $P(A|B)$, which denotes the probability of Event A, given that Event B has occurred. The vertical line between A and B denotes the "given" condition. (In this notation, the vertical line does not denote division).

For example, we might want to know the probability of a person getting a parking ticket given that a person did not put any money in a parking meter. Or a medical researcher might be interested in the probability of a patient developing heart disease given that the patient is a smoker.

If the occurrence of one event affects the probability of occurrence for another event, we say that the events are *dependent*; otherwise, the events are *independent*. **Dependent events** are events where the occurrence of one event affects the probability of occurrence of another event. **Independent events** are events where the probability of occurrence of one event is *not* affected by the occurrent of another event. The dependence of events has important implications in many fields such as marketing, engineering, psychology, and medicine.

EXAMPLE 3.18

Problem

Determine if the two events are dependent or independent:

1. Rolling a 3 on one roll of a die, rolling a 4 on a second roll of a die

2. Obtaining heads on one flip of a coin and obtaining tails on a second flip of a coin
3. Selecting five basketball players from a professional basketball team and a player's height is greater than 6 feet
4. Selecting an Ace from a deck of 52 cards, returning the card back to the original stack, and then selecting a King
5. Selecting an Ace from a deck of 52 cards, not returning the card back to the original stack, and then selecting a King

Solution

1. The result of one roll does not affect the result for the next roll, so these events are independent.
2. The results of one flip of the coin do not affect the results for any other flip of the coin, so these events are independent.
3. Typically, basketball players are tall individuals, and so they are more likely to have heights greater than 6 feet as opposed to the general public, so these events are dependent.
4. By selecting an Ace from a deck of 52 cards and then replacing the card, this restores the deck of cards to its original state, so the probability of selecting a King is not affected by the selection of the Ace. So these events are independent.
5. By selecting an Ace from a deck of 52 cards and then not replacing the card, this will result in only 51 cards remaining in the deck. Thus, the probability of selecting a King is affected by the selection of the Ace, so these events are dependent.

There are several ways to use conditional probabilities in data science applications.

Conditional probability can be defined as follows:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}, \text{ where } P(B) \neq 0$$

When assessing the conditional probability of $P(A|B)$, if the two events are independent, this indicates that Event A is not affected by the occurrence of Event B, so we can write that $P(A|B) = P(A)$ for independent events.

If we determine that the $P(A|B)$ is not equal to $P(A)$, this indicates that the events are dependent.

$P(A|B) = P(A)$ implies independent events, where $P(B) \neq 0$.

$P(A|B) \neq P(A)$ implies dependent events.

EXAMPLE 3.19

Problem

Table 3.2 shows the number of nursing degrees and non-nursing degrees at a university for a specific year, and the data is broken out by age groups. Calculate the probability that a randomly chosen graduate obtained a nursing degree, given that the graduate is in the age group of 23 and older.

Age Group	Nursing Degrees	Non-Nursing Degrees	Total
22 and under	1036	1287	2323
23 and older	986	932	1918
Total	2022	2219	4241

Table 3.2 Number of Nursing and Non-Nursing Degrees at a University by Age Group

Solution

Since we are given that the group of interest are those graduates in the age group of 23 and older, focus only on the second row in the table.

Looking only at the second row in the table, we are interested in the probability that a randomly chosen graduate obtained a nursing degree. The reduced sample space consists of 1,918 graduates, and 986 of them received nursing degrees. So the probability can be calculated as:

$$P(\text{nursing degree} \mid \text{age group of 23 and older}) = \frac{986}{1{,}918} = 0.514$$

Another method to analyze this example is to rewrite the conditional probability using the equation for $P(A \text{ and } B)$, as follows:

$$\begin{aligned} P(A \text{ and } B) &= P(A) \cdot P(B|A) \\ \text{rewrite as: } P(B|A) &= \frac{P(A \text{ and } B)}{P(A)} \end{aligned}$$

We can now use this equation to calculate the probability that a randomly chosen graduate obtained a nursing degree, given that the graduate is in the age group of 23 and older. The probability of A and B is the probability that a graduate received a nursing degree and is also in the age group of 23 and older. From the table, there are 986 graduates who earned a nursing degree and are also in the age group of 23 and older. Since this number of graduates is out of the total sample size of 4,241, we can write the probability of Events A and B as:

$$P(A \text{ and } B) = \frac{986}{4{,}241}$$

We can also calculate the probability that a graduate is in the age group of 23 and older. From the table, there are 1,918 graduates in this age group out of the total sample size of 4,241, so we can write the probability for Event A as:

$$P(A) = \frac{1{,}918}{4{,}241}$$

Next, we can substitute these probabilities into the formula for $P(B|A)$, as follows:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{\frac{986}{4{,}241}}{\frac{1{,}918}{4{,}241}} = \frac{986}{4{,}241} \cdot \frac{4{,}241}{1{,}918} = \frac{986}{1{,}918} = 0.514$$

Probability of *At Least One*

The probability of *at least one occurrence* of an event is often of interest in many data science applications. For example, a doctor might be interested to know the probability that at least one surgery to be performed this week will involve an infection of some type.

The phrase "at least one" implies the condition of one or more successes. From a sample space perspective, one or more successes is the complement of "no successes." Using the complement rule discussed earlier, we can write the following probability formula:

$$P(\text{at least one success}) = 1 - P(\text{no successes})$$

As an example, we can find the probability of rolling a die 3 times and obtaining at least one four on any of the rolls. This can be calculated by first finding the probability of not observing a four on any of the rolls and then subtracting this probability from 1. The probability of not observing a four on a roll of the die is 5/6. Thus, the probability of rolling a die 3 times and obtaining at least one four on any of the rolls is $1 - \left(\frac{5}{6}\right)^3 = 0.421$.

EXAMPLE 3.20

Problem

From past data, hospital administrators determine the probability that a knee surgery will be successful is 0.89.

1. During a certain day, the hospital schedules four knee surgeries to be performed. Calculate the probability that all four of these surgeries will be successful.
2. Calculate the probability that none of these knee surgeries will be successful.
3. Calculate the probability that at least one of the knee surgeries will be successful.

Solution

1. For all four surgeries to be successful, we can interpret that as the first surgery will be successful, and the second surgery will be successful, and the third surgery will be successful, and the fourth surgery will be successful. Since the probability of success for one knee surgery does not affect the probability of success for another knee surgery, we can assume these events are independent. Based on this, the probability that all four surgeries will be successful can be calculated using the probability formula for $P(A \text{ and } B)$ by multiplying the probabilities together:

$$\begin{aligned} P(A \text{ and } B \text{ and } C \text{ and } D) &= P(A) \cdot P(B) \cdot P(C) \cdot P(D) \\ &= 0.89 \cdot 0.89 \cdot 0.89 \cdot 0.89 = 0.627 \end{aligned}$$

   There is about a 63% chance that all four knee surgeries will be successful.

2. The probability that a knee surgery will be unsuccessful can be calculated using the complement rule. If the probability of a successful surgery is 0.89, then the probability that the surgery will be unsuccessful is 0.11:

$$P(A') = 1 - P(A) = 1 - 0.89 = 0.11$$

   Based on this, the probability that all four surgeries will be unsuccessful can be calculated using the probability formula for $P(A \text{ and } B)$ by multiplying the probabilities together:

$$\begin{aligned}(A' \text{ and } B' \text{ and } C' \text{ and } D') &= P(A') \cdot P(B') \cdot P(C') \cdot P(D') \\ &= 0.11 \cdot 0.11 \cdot 0.11 \cdot 0.11 = 0.00015\end{aligned}$$

Since this is a very small probability, it is very unlikely that none of the surgeries will be successful.

3. To calculate the probability that at least one of the knee surgeries will be successful, use the probability formula for "at least one," which is calculated as the complement of the event "none are successful."

$$\begin{aligned}P(\text{at least one success}) &= 1 - P(\text{no successes}) \\ P(\text{at least one success}) &= 1 - 0.00015 = 0.99985\end{aligned}$$

This indicates there is a very high probability that at least one of the knee surgeries will be successful.

Bayes' Theorem

Bayes' Theorem is a statistical technique that allows for the revision of probability estimates based on new information or evidence that allows for more accurate and efficient decision-making in uncertain situations. Bayes' Theorem is often used to help assess probabilities associated with medical diagnoses such as the probability a patient will develop cancer based on test screening results. This can be important in medical analysis to help assess the impact of a *false positive*, which is the scenario where the patient does not have the ailment but the screening test gives a false indication that the patient does have the ailment.

Bayes' Theorem allows the calculation of the conditional probability $P(A|B)$. There are several forms of Bayes' Theorem, as shown:

$$\begin{aligned}P(A|B) &= \frac{P(A) \cdot P(B|A)}{P(B)} \\ P(A|B) &= \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A') \cdot P(B|A')}\end{aligned}$$

EXAMPLE 3.21

Problem

Assume that a certain type of cancer affects 3% of the population. Call the event that a person has cancer "Event A," so:

$$P(A) = 0.03$$

A patient can undergo a screening test for this type of cancer. Assume the probability of a true positive from the screening test is 75%, which indicates that probability that a person has a positive test result given that they actually have cancer is 0.75. Also assume the probability of a false positive from the screening test is 15%, which indicates that probability that a person has a positive test result given that they do not have cancer is 0.15.

A medical researcher is interested in calculating the probability that a patient actually has cancer given that the screening test shows a positive result.

The researcher is interested in calculating $P(A|B)$, where Event A is the person actually has cancer and Event B is the event that the person shows a positive result in the screening test. Use Bayes' Theorem to calculate this conditional probability.

Solution

From the example, the following probabilities are known:

$$\begin{aligned} P(A) &= 0.03 \\ P(A') &= 0.97 \end{aligned}$$

The conditional probabilities can be interpreted as follows:

$$\begin{aligned} P(B|A) &= P(\text{positive test result} \mid \text{patient has cancer}) = 0.75 \\ P(B|A') &= P(\text{positive test result} \mid \text{patient does not have cancer}) = 0.15 \end{aligned}$$

Substituting these probabilities into the formula for Bayes' Theorem results in the following:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A') \cdot P(B|A')}$$

$$P(\text{ patient has cancer} \mid \text{positive test result}) = \frac{0.03 \cdot 0.75}{0.03 \cdot 0.75 + 0.97 \cdot 0.15} = 0.134$$

This result from Bayes' Theorem indicates that even if a patient receives a positive test result from the screening test, this does not imply a high likelihood that the patient has cancer. There is only a 13% chance that the patient has cancer given a positive test result from the screening test.

3.5 Discrete and Continuous Probability Distributions

Learning Outcomes

By the end of this section, you should be able to:

- 3.5.1 Describe fundamental aspects of probability distributions.
- 3.5.2 Apply discrete probability distributions including binomial and Poisson distributions.
- 3.5.3 Apply continuous probability distributions including exponential and normal distributions.
- 3.5.4 Use Python to apply various probability distributions for probability applications.

Probability distributions are used to model various scenarios to help with probability analysis and predictions, and they are used extensively to help formulate probability-based decisions. For example, if a doctor knows that the weights of newborn infants follow a normal (bell-shaped) distribution, the doctor can use this information to help identify potentially underweight newborn infants, which might indicate a medical condition warranting further investigation. Using a normal distribution, the doctor can calculate that only a small percentage of babies have weights below a certain threshold, which might prompt the doctor to further investigate the cause of the low weight. Or a medical researcher might be interested in the probability that a person will have high blood pressure or the probability that a person will have type O blood.

Overview of Probability Distributions

To begin our discussion of probability distributions, some terminology will be helpful:

- **Random variable**—a variable where a single numerical value is assigned to a specific outcome from an experiment. Typically the letter x is used to denote a random variable. For example, assign the numerical values 1, 2, 3, … 13 to the cards selected from a standard 52-card deck of Ace, 2, 3, … 10, Jack, Queen, King. Notice we cannot use "Jack" as the value of the random variable since by definition a random variable must be a numerical value.
- **Discrete random variable**—a random variable is considered discrete if there is a finite or countable number of values that the random variable can take on. (If there are infinitely many values, the number of values is countable if it possible to count them individually.) Typically, a discrete random variable is the result of a count of some kind. For example, if the random variable x represents the number of cars in a

parking lot, then the values that x can take on can only be whole numbers since it would not make sense to have $x = 15.37$ cars in the parking lot.

- **Continuous random variable**—a random variable is considered continuous if the value of the random variable can take on any value within an interval. Typically, a continuous random variable is the result of a measurement of some kind. For example, if the random variable x represents the weight of a bag of apples, then x can take on any value such as $x = 2.45734$ pounds of apples.

To summarize, the difference between discrete and continuous probability distributions has to do with the nature of the random variables they represent. **Discrete probability distributions** are associated with variables that take on a finite or countably infinite number of distinct values. **Continuous probability distributions** deal with random variables that can take on any value within a given range or interval. It is important to identify and distinguish between discrete and continuous random variables since different statistical methods are used to analyze each type.

EXAMPLE 3.22

Problem

A coin is flipped three times. Determine a possible random variable that can be assigned to represent the number of heads observed in this experiment.

Solution

One possible random variable assignment could be to let x count the number of heads observed in each possible outcome in the sample space. When flipping a coin three times, there are eight possible outcomes, and x will be the numerical count corresponding to the number of heads observed for each outcome. Notice that the possible values for the random variable x are 0, 1, 2 and 3, as shown in Table 3.3.

Result for Flip #1	Result for Flip #2	Result for Flip #3	Value of Random Variable x
Heads	Heads	Heads	3
Heads	Heads	Tails	2
Heads	Tails	Heads	2
Heads	Tails	Tails	1
Tails	Heads	Heads	2
Tails	Heads	Tails	1
Tails	Tails	Heads	1
Tails	Tails	Tails	0

Table 3.3 Result of Three Random Coin Flips

EXAMPLE 3.23

Problem

Identify the following random variables as either discrete or continuous random variables:

1. The amount of gas, in gallons, used to fill a gas tank
2. Number of children per household in a certain neighborhood

3. Number of text messages sent by a certain student during a particular day
4. Number of hurricanes affecting Florida in a given year
5. The amount of rain, in inches, in Detroit, Michigan, in a certain month

Solution

1. The number of gallons of gas used to fill a gas tank can take on any value, such as 12.3489, so this represents a continuous random variable.
2. The number of children per household in a certain neighborhood can only take on certain discrete values such as 0, 1, 2, 3, etc., so this represents a discrete random variable.
3. The number of text messages sent by a certain student during a particular day can only take on certain discrete values such as 26, 10, 17, etc., so this represents a discrete random variable.
4. The number of hurricanes affecting Florida in a given year can only take on certain values such as 0, 1, 2, 3, etc., so this represents a discrete random variable.
5. The number of inches of rain in Detroit, Michigan, in a certain month can take on any value, such as 2.0563, so this represents a continuous random variable.

Discrete Probability Distributions: Binomial and Poisson

Discrete random variables are of interest in many data science applications, and there are several probability distributions that apply to discrete random variables. In this chapter, we present the binomial distribution and the Poisson distribution, which are two commonly used probability distributions used to model discrete random variables for different types of events.

Binomial Distribution

The **binomial distribution** is used in applications where there are two possible outcomes for each trial in an experiment and the two possible outcomes can be considered as success or failure. For example, when a baseball player is at-bat, the player either gets a hit or does not get a hit. There are many applications of binomial experiments that occur in medicine, psychology, engineering, science, marketing, and other fields.

There are many statistical experiments where the results of each trial can be considered as either a success or a failure. For example, when flipping a coin, the two outcomes are heads or tails. When rolling a die, the two outcomes can be considered to be an even number appears on the face of the die or an odd number appears on the face of the die. When conducting a marketing study, a customer can be asked if they like or dislike a certain product. Note that the word “success” here does not necessarily imply a good outcome. For example, if a survey was conducted of adults and each adult was asked if they smoke, we can consider the answer “yes” to be a success and the answer “no” to be a failure. This means that the researcher can define success and failure in any way; however, the binomial distribution is applicable when there are only two outcomes in each trial of an experiment.

The requirements to identify a binomial experiment and apply the binomial distribution include:

- The experiment of interest is repeated for a fixed number of trials, and each trial is independent of other trials. For example, a market researcher might select a sample of 20 people to be surveyed where each respondent will reply with a “yes” or “no” answer. This experiment consists of 20 trials, and each person’s response to the survey question can be considered as independent of another person’s response.
- There are only two possible outcomes for each trial, which can be labeled as “success” or “failure.”
- The probability of success remains the same for each trial of the experiment. For example, from past data we know that 35% of people prefer vanilla as their favorite ice cream flavor. If a group of 15 individuals are surveyed to ask if vanilla is their favorite ice cream flavor, the probability of success for each trial will be 0.35.
- The random variable x will count the number of successes in the experiment. Notice that since x will count

the number of successes, this implies that x will be a discrete random variable. For example, if the researcher is counting the number of people in the group of 15 that respond to say vanilla is their favorite ice cream flavor, then x can take on values such as 3 or 7 or 12, but x could not equal 5.28 since x is counting the number of people.

When working with a binomial experiment, it is useful to identify two specific parameters in a binomial experiment:

1. The number of trials in the experiment. Label this as n.
2. The probability of success for each trial (which is a constant value). Label this as p.

We then count the number of successes of interest as the value of the discrete random variable. Label this as x.

EXAMPLE 3.24

Problem

A medical researcher is conducting a study related to a certain type of shoulder surgery. A sample of 20 patients who have recently undergone the surgery is selected, and the researcher wants to determine the probability that 18 of the 20 patients had a successful result from the surgery. From past data, the researcher knows that the probability of success for this type of surgery is 92%.

1. Does this experiment meet the requirements for a binomial experiment?
2. If so, identify the values of n, p, and x in the experiment.

Solution

1. This experiment does meet the requirements for a binomial experiment since the experiment will be repeated for 20 trials, and each response from a patient will be independent of other responses. Each reply from a patient will be one of two responses—the surgery was successful or the surgery was not successful. The probability of success remains the same for each trial at 92%. The random variable x can be used to count the number of patients who respond that the surgery was successful.
2. The number of trials is 20 since 20 patients are being surveyed, so $n = 20$.
   The probability of success for each surgery is 92%, so $p = 0.92$.
   The number of successes of interest is 18 since the researcher wants to determine the probability that 18 of the 20 patients had a successful result from the surgery, so $x = 18$.

When calculating the probability for x successes in a binomial experiment, a binomial probability formula can be used, but in many cases technology is used instead to streamline the calculations.

The **probability mass function (PMF)** for the binomial distribution describes the probability of getting exactly x successes in n independent Bernoulli trials, each with a probability p of success. The PMF is given by the formula:

$$P(X) = x = \binom{n}{k} p^x (1-p)^{n-x}$$

Where:
$P(X = x)$ is the probability that the random variable X takes on the value of exactly x successes
n is the number of trials in the experiment
p is the probability of success
x is the number of successes in the experiment

$\binom{n}{k}$ refers to the number of ways to choose x successes from $\binom{n}{k} = \frac{n!}{(n-x)!x!}$

Note: The notation $n!$ is read as n factorial and is a mathematical notation used to express the multiplication of $n\,(n-1)\,(n-2)\ldots(3)(2)(1)$. For example, $5! = (5)(4)(3)(2)(1) = 120$.

EXAMPLE 3.25

Problem

For the binomial experiment discussed in Example 3.24, calculate the probability that 18 out of the 20 patients will respond to indicate that the surgery was successful. Also, show a graph of the binomial distribution to show the probability distribution for all values of the random variable x.

Solution

In Example 3.24, the parameters of the binomial experiment are:
$n = 20$
$p = 0.92$
$x = 18$

Substituting these values into the binomial probability formula, the probability for 18 successes can be calculated as follows:

$$\begin{aligned} P(x) &= \frac{n!}{(n-x)!x!}p^x(1-p)^{n-x} \\ P(18) &= \frac{20!}{(20-18)!18!}0.92^{18}(1-0.92)^{20-18} \\ P(18) &= \frac{20!}{2!18!}0.92^{18}(0.18)^2 \\ P(18) &= 190(0.223)(0.032) \\ P(18) &= 0.271 \end{aligned}$$

Based on this result, the probability that 18 out of the 20 patients will respond to indicate that the surgery was successful is 0.271, or approximately 27%.

Figure 3.5 illustrates this binomial distribution, where the horizontal axis shows the values of the random variable x, and the vertical axis shows the binomial probability for each value of x. Note that values of x less than 14 are not shown on the graph since these corresponding probabilities are very close to zero.

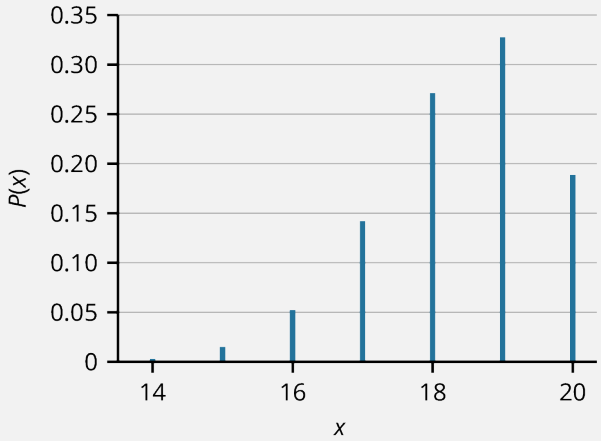

Figure 3.5 Graph of the Binomial Distribution for $n = 20$ and $p = 0.92$

Since these computations tend to be complicated and time-consuming, most data scientists will use technology (such as Python, R, Excel, or others) to calculate binomial probabilities.

Poisson Distribution

The goal of a binomial experiment is to calculate the probability of a certain number of successes in a specific number of trials. However, there are certain scenarios where a data scientist might be interested to know the probability of a certain number of occurrences for a random variable in a specific interval, such as an interval of time.

For example, a website developer might be interested in knowing the probability that a certain number of users visit a website per minute. Or a traffic engineer might be interested in calculating the probability of a certain number of accidents per month at a busy intersection.

The **Poisson distribution** is applied when counting the number of occurrences in a certain interval. The random variable then counts the number of occurrences in the interval.

A common application for the Poisson distribution is to model arrivals of customers for a queue, such as when there might be 6 customers per minute arriving at a checkout lane in the grocery store and the store manager wants to ensure that customers are serviced within a certain amount of time.

The Poisson distribution is a discrete probability distribution used in these types of situations where the interest is in a specific certain number of occurrences for a random variable in a certain interval such as time or area.

The Poisson distribution is used where the following conditions are met:

- The experiment is based on counting the number of occurrences in a specific interval where the interval could represent time, area, volume, etc.
- The number of occurrences in one specific interval is independent of the number of occurrences in a different interval.

Notice that when we count the number of occurrences that a random variable x occurs in a specific interval, this will represent a discrete random variable. For example, the count of the number of customers that arrive per hour to a queue for a bank teller might be 21 or 15, but the count could not be 13.32 since we are counting the number of customers and hence the random variable will be discrete.

To calculate the probability of x successes, the Poisson probability formula can be used, as follows:

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}, \text{ where } x = 0, 1, 2, \ldots$$

Where:
μ is the average or mean number of occurrences per interval
e is the constant 2.71828...

EXAMPLE 3.26

Problem

From past data, a traffic engineer determines the mean number of vehicles entering a parking garage is 7 per 10-minute period. Calculate the probability that the number of vehicles entering the garage is 9 in a certain 10-minute period. Also, show a graph of the Poisson distribution to show the probability distribution for various values of the random variable x.

Solution

This example represents a Poisson distribution in that the random variable x is based on the number of vehicles entering a parking garage per time interval (in this example, the time interval of interest is 10 minutes). Since the average is 7 vehicles per 10-minute interval, we label the mean μ as 7. Since the

engineer want to know the probability that 9 vehicles enter the garage in the same time period, the value of the random variable x is 9.

Thus, in this example, the parameters of the Poisson distribution are:

$$\mu = 7$$
$$x = 9$$

Substituting these values into the Poisson probability formula, the probability for 9 vehicles entering the garage in a 10-minute interval can be calculated as follows:

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$
$$P(9) = \frac{7^9 e^{-7}}{9!} = 0.101$$

Thus, there is about a 10% probability of 9 vehicles entering the garage in a 10-minute interval.

Figure 3.6 illustrates this Poisson distribution, where the horizontal axis shows the values of the random variable x and the vertical axis shows the Poisson probability for each value of x.

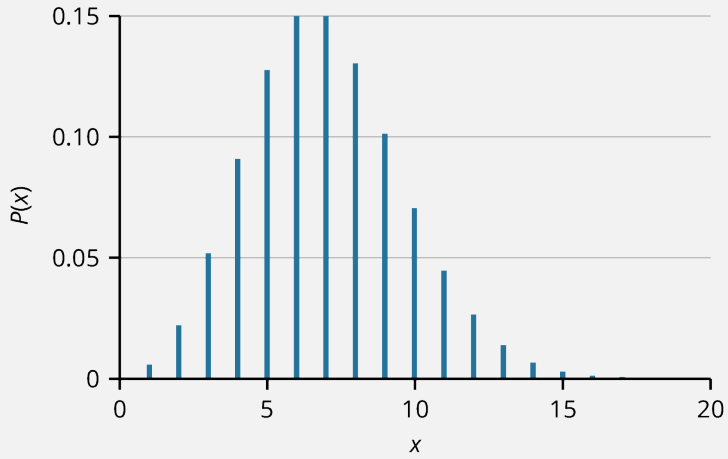

Figure 3.6 Poisson Distribution for $\mu = 7$

As with calculations involving the binomial distribution, data scientists will typically use technology to solve problems involving the Poisson distribution.

Normal Continuous Probability Distributions

Recall that a random variable is considered continuous if the value of the random variable can take on any of infinitely many values. We used the example about that if the random variable x represents the weight of a bag of apples, then x can take on any value such as $x = 2.45734$ pounds of apples.

Many probability distributions apply to continuous random variables. These distributions rely on determining the probability that the random variable falls within a distinct range of values, which can be calculated using a probability density function (PDF). The **probability density function (PDF)** calculates the corresponding area under the probability density curve to determine the probability that the random variable will fall within this specific range of values. For example, to determine the probability that a salary falls between $50,000 and $70,000, we can calculate the area under the probability density function between these two salaries.

Note that the total area under the probability density function will always equal 1. The probability that a continuous random variable takes on a specific value x is 0, so we will always calculate the probability for a random variable falling within some interval of values.

In this section, we will examine an important continuous probability distribution that relies on the probability density function, namely the *normal distribution*. Many variables, such as heights, weights, salaries, and blood pressure measurements, follow a normal distribution, making it especially important in statistical analysis. In addition, the normal distribution forms the basis for more advanced statistical analysis such as confidence intervals and hypothesis testing, which are discussed in Inferential Statistics and Regression Analysis.

The **normal distribution** is a continuous probability distribution that is symmetrical and bell-shaped. It is used when the frequency of data values decreases with data values above and below the mean. The normal distribution has applications in many fields including engineering, science, finance, medicine, marketing, and psychology.

The normal distribution has two parameters: the mean, μ, and the standard deviation, σ. The mean represents the center of the distribution, and the standard deviation measures the spread, or dispersion, of the distribution. The variable x represents the realization, or observed value, of the random variable X that follows a normal distribution.

The typical notation used to indicate that a random variable follows a normal distribution is as follows: $X \sim N(\mu, \sigma)$ (see Figure 3.7). For example, the notation $X \sim N(5.2, 3.7)$ indicates that the random variable follows a normal distribution with mean of 5.2 and standard deviation of 3.7.

A normal distribution with mean of 0 and standard deviation of 1 is called the **standard normal distribution** and can be notated as $X \sim N(0, 1)$. Any normal distribution can be standardized by converting its values to z-scores. Recall that a z-score tells you how many standard deviations from the mean there are for a given measurement.

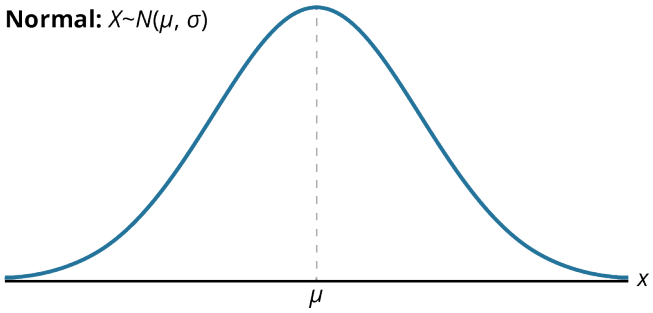

Figure 3.7 Graph of the Normal (Bell-Shaped) Distribution

The curve in Figure 3.7 is symmetric on either side of a vertical line drawn through the mean, μ. The mean is the same as the median, which is the same as the mode, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Because the area under the curve must equal 1, a change in the standard deviation, σ, causes a change in the shape of the normal curve; the curve becomes fatter and wider or skinnier and taller depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions.

To determine probabilities associated with the normal distribution, we find specific areas under the normal curve. There are several methods for finding this area under the normal curve, and we typically use some form of technology. Python, Excel, and R all provide built-in functions for calculating areas under the normal curve.

EXAMPLE 3.27

Problem

Suppose that at a software company, the mean employee salary is \$60,000 with a standard deviation of \$7,500. Assume salaries at this company follow a normal distribution. Use Python to calculate the

probability that a random employee earns more than $68,000.

Solution

A normal curve can be drawn to represent this scenario, in which the mean of $60,000 would be plotted on the horizontal axis, corresponding to the peak of the curve. Then, to find the probability that an employee earns more than $68,000, calculate the area under the normal curve to the right of the data value $68,000.

Figure 3.8 illustrates the area under the normal curve to the right of a salary of $68,000 as the shaded-in region.

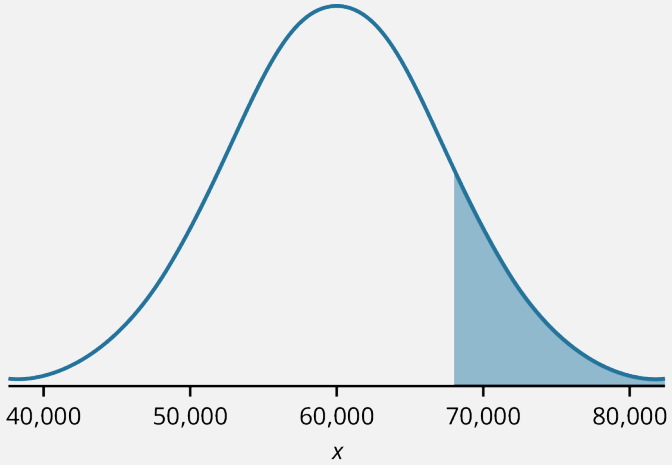

Figure 3.8 Bell-Shaped Distribution for Example 3.27. The shaded region under the normal curve corresponds to the probability that an employee earns more than $68,000.

To find the actual area under the curve, a Python command can be used to find the area under the normal probability density curve to the right of the data value of $68,000. See Using Python with Probability Distributions for the specific Python program and results. The resulting probability is calculated as 0.143.

Thus, there is a probability of about 14% that a random employee has a salary greater than $75,000.

The **empirical rule** is a method for determining approximate areas under the normal curve for measurements that fall within one, two, and three standard deviations from the mean for the normal (bell-shaped) distribution. (See Figure 3.9).

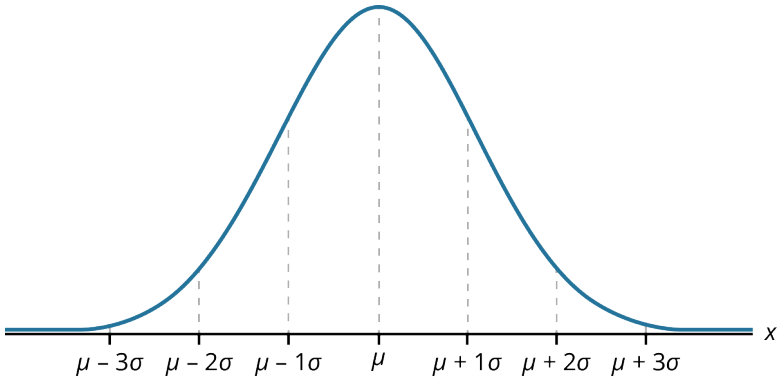

Figure 3.9 Normal Distribution Showing Mean and Increments of Standard Deviation

If x is a continuous random variable and has a normal distribution with mean μ and standard deviation σ, then the empirical rule states that:

- About 68% of the x-values lie between -1σ and $+1\sigma$ units from the mean μ (within one standard deviation

of the mean).
- About 95% of the x-values lie between -2σ and $+2\sigma$ units from the mean μ (within two standard deviations of the mean).
- About 99.7% of the x-values lie between -3σ and $+3\sigma$ units from the mean μ (within three standard deviations of the mean). *Notice that almost all the x-values lie within three standard deviations of the mean.*
- The z-scores for $+1\sigma$ and -1σ are $+1$ and -1, respectively.
- The z-scores for $+2\sigma$ and -2σ are $+2$ and -2, respectively.
- The z-scores for $+3\sigma$ and -3σ are $+3$ and -3, respectively.

EXAMPLE 3.28

Problem

An automotive designer is interested in designing automotive seats to accommodate the heights for about 95% of customers. Assume the heights of adults follow a normal distribution with mean of 68 inches and standard deviation of 3 inches. For what range of heights should the designer model the car seats to accommodate 95% of drivers?

Solution

According to the empirical rule, the area under the normal curve within two standard deviations of the mean is 95%. Thus, the designer should design the seats to accommodate heights that are two standard deviations away from the mean. The lower bound of heights would be $68 - 2(3)$ inches, and the upper bound of heights would be $68 + 2(3)$ inches. Thus, the car seats should be designed to accommodate driver heights between 62 and 74 inches.

EXPLORING FURTHER

Statistical Applets to Explore Statistical Concepts

Applets are very useful tools to help visualize statistical concepts in action. Many applets can simulate statistical concepts such as probabilities for the normal distribution, use of the empirical rule, creating box plots, etc.

Visit the Utah State University applet website (https://openstax.org/r/usu) and experiment with various statistical tools.

Using Python with Probability Distributions

Python provides a number of built-in functions for calculating probabilities associated with both discrete and continuous probability distributions such as binomial distribution and the normal distribution. These functions are part of a library called scipy.stats (https://openstax.org/r/scipy).

Here are a few of these probability density functions available within Python:

- `binom()`—calculate probabilities associated with the binomial distribution
- `poisson()`—calculate probabilities associated with the Poisson distribution
- `expon()`—calculate probabilities associated with the exponential distribution
- `norm()`—calculate probabilities associated with the normal distribution

To import these probability density functions within Python, use the import command. For example, to import

the `binom()` function use the following command:

```
from scipy.stats import binom
```

Using Python with the Binomial Distribution

The `binom()` function in Python allows calculations of binomial probabilities. The probability mass function for the binomial distribution within Python is referred to as `binom.pmf()`.

The syntax for using this function is `binom.pmf(x, n, p)`

Where:
n is the number of trials in the experiment
p is the probability of success
x is the number of successes in the experiment

Consider the previous Example 3.24 worked out using the Python `binom.pmf()` function. A medical researcher is conducting a study related to a certain type of shoulder surgery. A sample of 20 patients who have recently undergone the surgery is selected, and the researcher wants to determine the probability that 18 of the 20 patients had a successful result from the surgery. From past data, the researcher knows that the probability of success for this type of surgery is 92%. Round your answer to 3 decimal places.

In this example:
n is the number of trials in the experiment $= 20$
p is the probability of success $= 0.92$
x is the number of successes in the experiment $= 18$

The corresponding function in Python is written as:

```
binom.pmf (18, 20, 0.92)
```

The `round()` function is then used to round the probability result to 3 decimal places.

Here is the input and output of this Python program:

PYTHON CODE

```
# import the binom function from the scipy.stats library
from scipy.stats import binom

# define parameters x, n, and p:
x = 18
n = 20
p = 0.92

# use binom.pmf() function to calculate binomial probability
# use round() function to round answer to 3 decimal places
round (binom.pmf(x, n, p), 3)
```

The resulting output will look like this:

```
0.271
```

Using Python with the Normal Distribution

The `norm()` function in Python allows calculations of normal probabilities. The probability density function is sometimes called the cumulative density function, and so this is referred to as `norm.cdf()` within Python. The `norm.cdf()` function returns the area under the normal probability density function to the left of a specified measurement.

The syntax for using this function is
`norm.cdf(x, mean, standard_deviation)`

Where:
x is the measurement of interest
`mean` is the mean of the normal distribution
`standard_deviation` is the standard deviation of the normal distribution

Let's work out the previous Example 3.27 using the Python `norm.cdf()` function.

Suppose that at a software company, the mean employee salary is $60,000 with a standard deviation of $7,500. Use Python to calculate the probability that a random employee earns more than $68,000.

In this example:
x is the measurement of interest $= 68{,}000$
mean is the mean of the normal distribution $= 60{,}000$
standard deviation is the standard deviation of the normal distribution $= 7{,}500$

The corresponding function in Python is written as:
`norm.cdf(68000, 60000, 7500)`

The `round()` function is then used to round the probability result to 3 decimal places.

Notice that since this example asks to find the area to the right of a salary of $68,000, we can first find the area to the left using the `norm.cdf()` function and subtract this area from 1 to then calculate the desired area to the right.

Here is the input and output of the Python program:

PYTHON CODE

```
# import the norm function from the scipy.stats library
from scipy.stats import norm
# define parameters x, mean and standard_deviation:
x = 68000
mean = 60000
standard_deviation = 7500
# use norm.cdf() function to calculate normal probability - note this is
# the area to the left
# subtract this result from 1 to obtain area to the right of the x-value
# use round() function to round answer to 3 decimal places
round (1 - norm.cdf(x, mean, standard_deviation), 3)
```

The resulting output will look like this:

```
0.143
```