|  |
| --- |
| IUBH |
| Applied Statistics |
| MMET01-01\_E |

# Learning Objectives

The world we live in contains large amounts of data, which are collected every second, minute, and hour. Data from our phones, cars, computers, and even our own bodies are con- stantly being uploaded to servers. These data have insights embedded within them that are challenge for humans to extract without the special tools available to us through the study of statistics. Statistics is a special branch of mathematics that addresses the collection, analy- sis, and interpretation of numerical data. In the course Applied Statistics, we will study the discipline of statistics and recognize its special role and importance in practical decision- making processes in business. Speciﬁcally, we will learn to understand the relevance of data to answer empirical questions, apply statistical methods in the overall context of concrete problems, and solve statistical problems by using special statistical software.

Statistical analysis (also called statistics) is an important tool used by business managers to make decisions in uncertain conditions. A statistic is a single number used to explain a set of data. This course will teach students to understand and execute these tools to enable better business decisions.

# Unit 1 – Basics

### Study Goals

On completion of this unit, you will be able to …

… deﬁne and compute descriptive statistics.

… extract insights from a sample using inferential statistics.

… compute and interpret probabilities.

# 1. Basics

## Case Study

The company, TechnoBank, is a ﬁnancial services company that has faced increased competition over the last few years and desires to gather data on its customers to bet- ter understand customer sentiments and loyalty. The marketing team, under the lead- ership of the chief marketing ofﬁcer (CMO) Kristin Stark, decides to design a question- naire and conduct a sample of current and lapsed customers. Kristen asks the market research lead, Stefan Rider, to lead the project. Stefan outlines the process of question- naire design and sampling from the TechnoBank online banking access application. He plans to poll customers in a systematic manner (one out of every ﬁve), conduct descriptive statistics to describe measures of loyalty and customer satisfaction, and then use these data to calculate the probability of customer attrition. Kristen is requesting an update in the next week and Stefan is contemplating four questions:

1. What sample size should be chosen?
2. Will this type of sampling method be an effective estimate for the entire popula- tion?
3. What measures of central tendency, dispersion, and graphics would be effective in telling a story with the data?
4. How should the right questions be asked to gather the proper loyalty data, and how should the probability of customer attrition then be calculated?

These are just some of the statistical decisions we will be exploring in this unit to ena- ble important business decisions like the ones TechnoBank is working on.

## 1.1 Descriptive Statistics

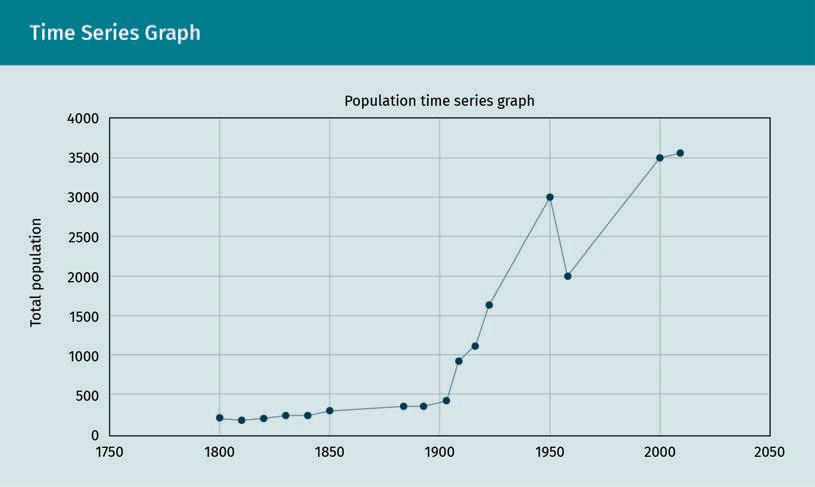
Descriptive statistics represent the foundation of any statistical analysis. When we want to begin to understand the data from a data ﬁle, the ﬁrst step is to describe them. Descriptive statistics are the primary tool used to accomplish this objective. Without the help of descriptive statistics, humans cannot make sense of more than a few rows of data organized in a column in a spreadsheet. Descriptive statistics allow researchers to take all of the individual observations of a variable and provide a summary in terms of the most likely values and how different they are to one another.

Descriptive statistics are often reported in conjunction with one another, since any sin- gle number or univariate statistic is not usually sufﬁcient to fully explain the character- istics of the variables in the dataset. For example, it is misleading to report the mean as an estimator of the population if the data are highly skewed. Moreover, there are some phenomena that are bimodal and should, therefore, be reported as two distinct distributions. Finally, some variables reported in conjunction with one another can give the researcher an idea of what is happening in the distribution (i.e., the mean and the median signiﬁcantly varying from one another could indicate a non-normal distribution). Descriptive statistics are often thought of as being a lower form of data analysis on a maturity curve representing the competitive advantages of analytics. This is a mis- nomer, as descriptive statistics are needed to support many higher-level analyses.

### Data Terminology

The ﬁrst discussion we need to have when beginning a study of statistics is an overview of data terminology. When discussing data, we need to deﬁne a variable that is a char- acteristic of a subject or object that varies. Examples of variables include age, gender, and income. These data elements will vary in relation to the individuals that are being measured or proﬁled. The term “data” is plural and refers to of all or part of a collec- tion of observations from the variables, while a dataset is all of the values of one or more variables collected for research purposes. For many, the initial exposure to varia- bles, data, and datasets occurs while using Excel spreadsheets.

Next, we will cover the types of data. Categorical data, also known as qualitative data, are data described in words, rather than numerically. An example of a categorical varia- ble is car colors. Numeric data come from counting measurements or a type of mathe- matical operation. Examples of numeric variables include income and time of day. The two major types of numeric data are discrete and continuous variables. Discrete varia- bles are those which can be counted, such as the number of customers entering a store. Continuous data are those variables which cannot be counted, such as weight and time, which can both be measured ad inﬁnitum. Another type of data worth men- tioning, especially in relation to business analytics, is time series data. Time series data are measured consecutively over a period of time. The ﬁgure below is a graph repre- senting time-series data.



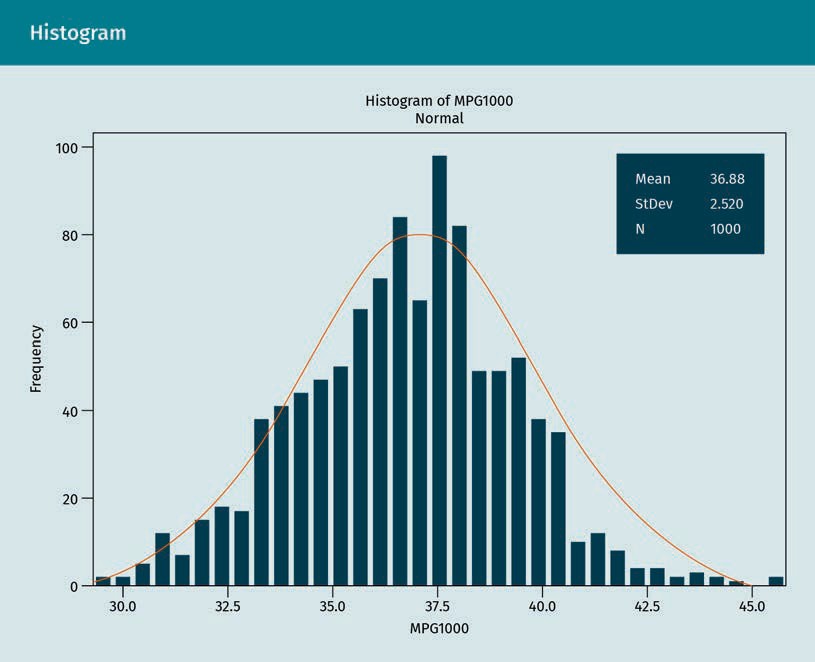
### Levels of Measurement

The levels of measurement for data are important in choosing the right statistical tech- niques to incorporate. The four levels of measurement for data are nominal, ordinal, interval, and ratio. Nominal data, derived from the Latin nomen or name, identify a cat- egory (Doane, 2016). Categorical data, such as hair color, are measured at the nominal level. Next, we have the ordinal level of measurement. Ordinal data, such as credit rat- ings of companies, are ranked data, but the difference between observations is not meaningful outside of the ranks. Then we arrive at the interval level of measurement, where the difference between scale points has meaning. However, there is no natural zero, and the difference between the scale points cannot be used in a ratio calculation because they are not fully known. Fahrenheit and Celsius are examples of measure- ments or variables at the interval level.

The ﬁnal level of measurement we will be discussing is the ratio level of measurement. Variables at the ratio level of measurement have a natural zero, which can be meas- ured and makes the use of ratios possible with the data. The temperature measure, Kelvin, is an example of a variable at the ratio level of measurement (as opposed to the aforementioned Fahrenheit or Celsius measures) because Kelvin measures the kinetic energy of small particles using the law of thermodynamics, and therefore has a natural zero. Ratio and interval levels of measurement are most suitable for statistical techni- ques. Income is an example of a variable at the ratio level of measurement, although some would still argue that the presence of credit could take a person or corporate entity below zero. Nonetheless, there would still be a point at which credit is no longer obtainable, thus creating a natural zero point (Ahmad, 2016; Doane, 2016).

### Measures of Center

Measures of center depend on the fact that many distributions in our world have some form of central tendency, meaning most of the observations are likely to be close to the center of a distribution rather than being outliers. A histogram is typically used to measure the shape of a distribution. The example histogram shown below portrays this tendency for a typical variable.



Before we delve into the measures of center, some knowledge of basic statistical nota- tion is necessary. Please review the list of statistical notations below (Byjus Classes, 2021). Note that this list does not contain all possible statistical notations, but provides the relevant notations for this course.

|  |  |  |  |
| --- | --- | --- | --- |
| *P*(*A*) = | | Probability function | |
| P(A | B) = | | Conditional probability function | |
| P(A ∪ B) = | | Probability of events union | |
| µ*=* | | Population mean | |
| var(X) = | | Variance | |
| E(X | Y) = | | Conditional expectation | |
| std(X) = | | Standard deviation | |
| σ*2 =* | | Variance | |
| x˜ = | | Median | |
| σ*X =* | | Standard deviation | |
| corr(X,Y) = | | Correlation | |
| cov(X,Y) = | | Covariance | |
| ρX,Y *=* | | Correlation | |
| Mo = | | Mode | |
| Md = | | Sample median | |
| MR *=* | | Mid-range | |
| Q2 = | | Median or second quartile | |
| Q1 = | | Lower or ﬁrst quartile | |
| x = | | Sample mean | |
| Q3 = | | Upper or third quartile | |
| s = | | Sample standard deviation | |
| s2 = | | Sample variance | |
| X ~ = | | Distribution of X | |
| zx = | | Standard score | |

### Mean

The **mean** is the most familiar and versatile measure of center. There are formulas for the population and sample mean. However, since we are most concerned with calculat- ing the mean from a sample, we use equation (1.1) to calculate the mean of a popula- tion (Ahmad, 2016; Doane, 2016):

**Mean**

If the data are nor- mally distributed, the mean should always be chosen as the measure of cen- ter because it can be used later in more powerful statistical techniques.

(1.1)

We can also calculate the mean using the Excel function =AVERAGE(data), where data refers to a column variable containing the actual data. It should be noted that the mean is only suitable for variables at the interval or ratio levels of measurement (Microsoft, n.d.). The arithmetic mean, also known as the “average,” is the balancing point if we were to weigh all of the variables on a scale. It is the balancing point because the sum of the mean distances to each of the data points in a distribution is always equal to zero. It should be noted that the mean will generally turn out to be less accurate than a measure of the typical value in the dataset if there are outliers or extreme measures in the data. For this reason, the mean should be presented with the **standard deviation**.

**Standard deviation**

The standard deviation measures the dispersion of observations around the mean.

### Median

The median represents the ﬁftieth percentile, midpoint, or centermost value of the sor- ted data variable. If the data have an uneven number of observations, then the median is calculated by taking the average of the two middle observations. The median should be used in place of the mean in the case of extreme values in the dataset. However, it should be noted that the median lacks some of the useful mathematical properties that the mean has, making it less versatile (Doane, 2016).

### Mode

The mode is the most frequent value in the data. If the data tend to be centrally dis- tributed, then the mean, median, and mode may be similar. However, the mode may also not be present at all in a dataset. Instead, a dataset may have several frequent variables, making it bimodal. To identify the mode, one has to observe the frequency of each value of a variable (for example, using the Excel function = MODE.SNGL(data variable)) (Microsoft, n.d.). The mode is frequently used to describe central tendency in categorical data or variables at the nominal level of measurement (Doane, 2016). There are other measures of center, but the mean, median, and mode are the most fre- quently used. Please note that, in a symmetrical distribution where the observations cluster around the center (normally distributed), the mean, median, and mode can all be very similar.

### Measures of Dispersion

The shape of a distribution can be observed through the use of a histogram. One can also compare the mean and median to visualize this. If the data are skewed right, it signiﬁes that the mean exceeds the median. Alternately, if the data are skewed left, this is an indication that the mean is below the median. When the mean, median, and even the mode are similar, then the distribution is considered symmetric. The difference between the data from the state of being symmetric is known as skewness.

Measures of dispersion are used to measure how the observations cluster around the mean. This measurement is known as the variation in the data. A simple measure of dispersion is the range, which is the highest value of the data subtracted from the low- est value, and is calculated using equation (1.2) (Doane, 2016):

Range = largest data value - smallest data value

(1.2)

In Excel, the range can be calculated using the function = MAX(data variable)- MIN(data variable) (Microsoft, n.d.). The most popular and versatile measure of dis- persion is the standard deviation (Ahmad, 2016). The standard deviation measures the average distance between the data variables from the mean, and is calculated using equation (1.3):

(1.3)

In Excel, the standard deviation can be calculate using the function = STD.S(data variable) (Microsoft, n.d.). Typically, the standard deviation is presented alongside the mean to further justify the usefulness of the mean in describing the data. The mean is not as useful as a descriptive statistic if the variable has a large standard deviation (Doane, 2016).

### Additional Descriptive Statistics

**Boxplot**

The whiskers on a box and whisker plot represent the quartiles of the distribution. The outiers are typically reperestented as asterisks in the outer region.

While measures of center and measures of dispersion are the most important general descriptive statistics used by researchers, other relevant measures include frequency (percent, count, frequency) and measures of position (quartile ranks and percentile ranks). A **boxplot** is often used to depict the median, quartiles, and even the outliers in the data.

## 1.2 Inferential Statistics

**Inferential statistics** are primarily concerned with inferring information about a popu- lation from a given sample. In many cases, it is either impractical or impossible to gather the desired information from an entire population. One common case is politi- cal polls, which, rather than doing a census of the entire population, provide a fairly accurate read within a few percentage points from a sample of several thousand peo- ple (provided appropriate probability sampling methods are conducted). This is a huge cost saving when compared to taking a sample of the entire population. If you were tasked with counting how many salmon currently live in the Columbia River, this would be an impossible task. However, with the correct planning and execution, this could be done in a fairly straightforward manner using sampling. In some cases, like conducting a census of the entire population of a country, which is often mandated by law, a sur- vey can be even more accurate. Trying to obtain the information from every individual residing in a country is challenging, despite having the proper resources to be able to conduct the census. The types of analysis one would conduct using inferential statistics include estimating parameters, testing hypotheses, and regression and trend analysis. There are also many aspects of quality control that make use of inferential statistics.

**Inferential statistics**

The process of using inferential statistics is sometimes referred to as the practice of statistical inference. Statistical inference depends heavily on sampling theory and central tendency.

In order to perform inferential statistics, it is important to sample the data from a pop- ulation in the proper manner. This ensures that the inferential statistics are unbiased estimators of the population. Random sampling methods include simple random sam- pling, where everyone in a population has an equal chance of being selected. Simple random sampling is a subset of **probability sampling** methods, where every element of the population (that is greater than zero) has a chance of being selected. Types of probability sampling, in addition to simple random sampling, include systematic, clus- ter, and stratiﬁed sampling. Since these methods all give elements of the population a chance of being selected, they are considered to be the most desirable sampling meth- ods for estimating a population using inferential statistics (Knapp, 1980). Other sam- pling methods in which all members or elements of a population do not have a chance of being selected include convenience, volunteer, and snowball sampling. While meth- ods like convenience sampling, which draw samples based on convenient segments of a population, tend to be popular in practice, they may create severe bias (Doane, 2016; Knapp, 1980).

**Probability sampling**

It is worth noting that, for probability sampling, researchers must use calibration techniques to make sure their resulting insights are reﬂective of the population proportions.

Once a sampling method has been chosen, the next step is to collect data from a sam- ple. One of the primary methods of collecting data from a sample is through a ques- tionnaire. Questionnaires can be custom-developed to collect data on virtually any research topic, but researchers must take extra precautions to ensure that the insights derived from the questionnaire are both valid and reliable with regard to the topic of interest.

### Estimating a Popular Parameter from a Statistic

Once we have our sample, we are concerned with estimating population parameters from the derived statistics. A statistic is a measure that is derived from a sample, and a parameter is a measure that describes the entire population, from which the sample was selected. All samples have random error, therefore, the difference between a statis- tic and parameter is known as the sampling error. Since we usually have no way of knowing the population parameter, we can use inferential statistics to estimate the parameters using hypothetical distribution, which takes sampling error into account.

A point estimate is a single value estimate of a parameter (a mean is an example of a point estimate). An interval estimate is a range of values within which the parameter most likely resides. While point estimates are easy to understand, they don’t give us much information on the variability of the parameter. A conﬁdence interval is one of the most popular interval estimates that provides a solution to this problem. Every conﬁdence interval includes a conﬁdence level derived from a theoretical distribution, which describes the probability of the interval containing the parameter estimate under repeated sampling conditions.

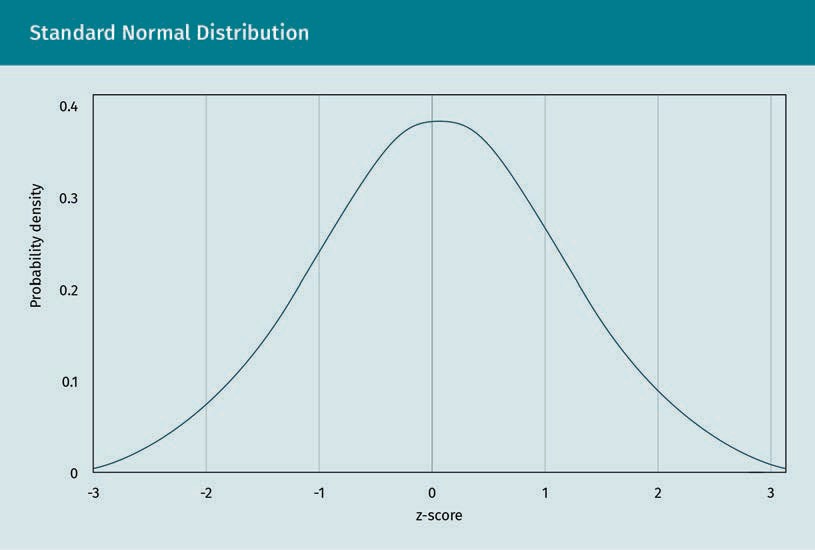
A 95 percent conﬁdence interval means that, with an inﬁnite number of samples, the interval will contain the true population parameter 95 times out of 100. The calculation for a conﬁdence interval can be found in equation (1.4). Hypothesis testing, correlation analysis. and regression analysis are additional types of inferential statistics.

(1.4)

|  |  |
| --- | --- |
| CI = | Conﬁdence interval |
| x = | Sample mean |
| z = | Conﬁdence level value |
| s = | Sample standard deviation |
| n = | Sample size |

where:

A z-value, or z-score, describes the position of a raw score in terms of its distance from the mean when measured in standard units of deviation (standard deviation) in a nor- mal distribution (Doane, 2016; Knapp, 1980). The ﬁgure below depicts the standard nor- mal distribution.



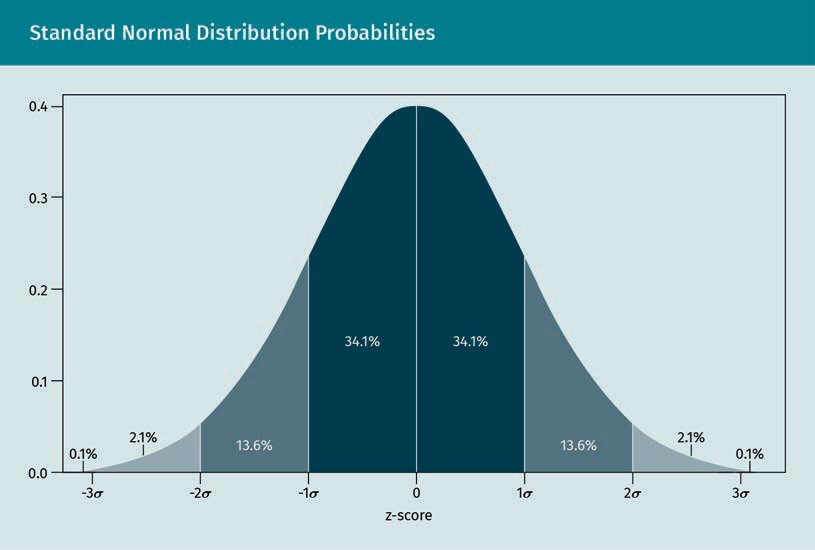
The z-score is positive for all values lying above the mean and negative for all values lying below the mean. The standard normal distribution shown in the figure above is a bell-shaped (normal) distribution which has a mean of zero and a standard deviation equal to one. The formula to calculate a z-score is z = (x-μ)/σ, where x is the raw score, μ is the population mean, and σ is the population standard deviation. Of course, since we rarely know the population mean and standard deviation, we can substitute the sample mean (x̄) and sample standard deviation (s) in the formula as unbiased estimates of the population parameters (Doane, 2016).

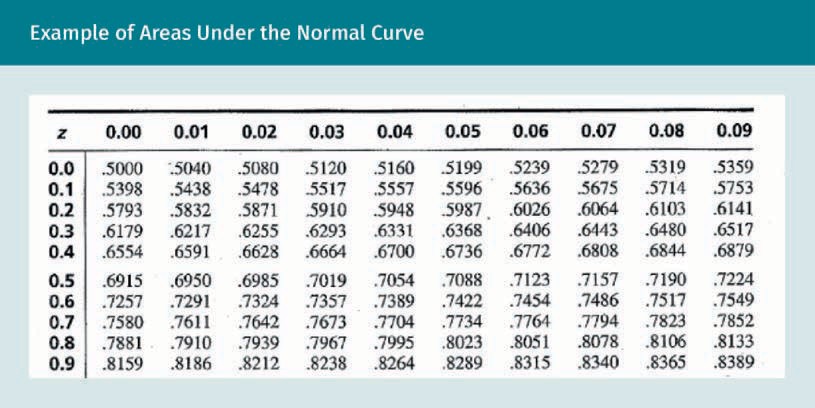
### Z-Calculation

The calculation for the z-score *z* is depicted below, where x represents the score, μ represents the mean, and sigma σ represents the standard deviation:

(1.5)

Once you calculate a z-score, it can be looked up in a z-table, part of which is depicted following the ﬁgure of the standard normal distribution below. For access to the full table see the literature by Farber and Larson (2017). The z-table represents the proba- bilities under the standard normal distribution, which enables researchers to calculate the probability of randomly obtaining a score from the sample distribution. For exam- ple, there is a 95.45 percent probability of randomly selecting a score between -2 and +2 standard deviations from the mean, as shown in the ﬁgure below (Doane, 2016; Knapp, 1980).





## 1.3 Probability Calculation

Archeological digs have discovered that ancient societies played some sort of probabil- ity games with primitive dice-like objects made from animal bones (Britannica, n.d.). However, it wasn’t until people created organized gambling that the serious study of probability was commenced. There are still many outstanding questions concerning what true probability means. Classical physics would state that true probability exists because we have unknown inputs. However, quantum physics, which dominates our lives today in terms of communications and computing, paints a different picture that there are, in fact, truly probabilistic phenomenon as long as we look at the world from a subatomic perspective (Fogarty, 2017). So, how do we think about probabilities on a day-to-day basis? Firstly, let’s examine the two methods of calculating probabilities. The ﬁrst is the classical probability method. With classical probabilities, every event has an equal likelihood of occurring. For example, rolling a die deemed to be fair should have an equal likelihood of rolling a 1, 2, 3, 4, 5, or 6. The same holds true for coin ﬂipping, where one has an equal chance of producing heads or tails, and choosing any one card from a standard deck of cards produces a has of 1/52 (Di Paola et al., 2018; Doane, 2016).

### Formulas for Classical Probability

Classical probability, also referred to as theoretical probability, is the expected proba- bility that a simple event will happen (i.e., rolling a die). This is the number of times the event can occur (one roll of the die) divided by the number of possible events (six sides). P(A) stands for “probability of event A,” with event A being whatever event that is being assessed, e.g., rolling a die and getting a 6. f is the frequency, or number of possible times the intended outcome could occur, and N is the number of times the event could happen.

For example, the odds of rolling a three on a fair die are one out of six, or 1/6. That’s one possible outcome divided by the number of possible outcomes (in this case, six). The odds of winning the Powerball lottery are 1/292.2 million (Amadeo, 2021). The “1” is the number of times the event can happen (i.e., you win the lottery), divided by the number of possible number combinations (about 292,000,000) on tickets sold. If you want to calculate the Mega-Millions, this would be 1/302.5 million (Amadeo, 2021).

### Empirical Probability

**Empirical or experimental probability** is different from classical probability in that it doesn’t assume a hypothetical probability based on the number of trials and possible outcomes, but instead seeks to calculate the actual probability of events occurring in an experiment (Doane, 2016). An interesting example of this is a statistician soldier from the US who was a prisoner of war during the Vietnam Conﬂict and recorded the results from hundreds of thousands of repeated coin ﬂips done during his time in the prison cell. His results showed that the results were slightly biased toward heads. Clas- sical probability would have said the chances were 50/50, but empirical probability calculations show that it would be better to always choose heads (Di Paola et al., 2018; Doane, 2016). Mathematically, the formula for empirical probability is:

**Empirical probability**

Sometimes, even the most widely accep- ted classical proba- bilities (e.g., coin toss probability) can be upended by empirical probabili-

ties

Empirical probability = (# of times an event occurs / total number of trials)

(1.6)

The experimental or empirical probability of an event is based on what has actually happened in a particular experiment, while the classical or theoretical probability attempts to forecast what will happen on the basis of total number of possible out- comes. We expect the experimental and theoretical probabilities to converge as the number of trials in an experiment increases (Di Paola et al., 2018; Doane, 2016).

### Subjective Probability

**Subjective probability** applies in the absence of a random repeatable experiment. Sub- jective probability uses informed judgment to predict the likelihood of an event. Inter- estingly, we use subjective probabilities in our everyday lives. Some real-life examples where humans may use subjective probability are

**Subjective probability**

The human mind has evolved to help us calculate subjec-tive probabilities. Moreover,many business decisions made using subjec-tive probabilities should be validated using empirical probabilities.

* predicting the outcome of a job interview,
* placing a bid accepted on a house, and
* calculating the risk of not wearing a seat belt.

### The Seven Basic Rules of Probability

The basic rules of probability are both logical and intuitive. Sometimes, probability problems become complicated, but can still be solved as long as these rules are applied in a consistent manner. In order to apply most of these rules, both logic and counting skills are necessary (Di Paola et al., 2018; Doane, 2016).

### Simple Probability (Rules 1—4)

#### Probability rule one

The ﬁrst basic rule of probability is that all probabilities are represented by a number between zero and one. The probability calculation is then incorrect if it is less than zero or greater than one (Di Paola et al., 2018; Doane, 2016):

For an event A, 0 ≤ P(A) ≤ 1.

(1.7)

#### Probability rule two

The second basic rule of probability is that the combined outcomes of all probabilities must be equal to one (Di Paola et al., 2018; Doane, 2016):

Sum of all outcomes = 1

(1.8)

#### Probability rule three

The third rule, also called the complement rule, is that the probability of an event not occuring is given by one minus the probability that an event will occur (Di Paola et al., 2018; Doane, 2016):

Sum of all outcomes = 1

*P*(*A*¯¯¯)=1−*P*(*A*)??

(1.9)

#### Probability rule four

The fourth basic rule of probability is also known as the addition rule. This rule states the probability of two disjoint events is the sum of the probability that either will hap- pen (Di Paola et al., 2018; Doane, 2016). If events A and B are disjoint events, then

P(A or B) = P(A) + P(B)

(1.10)

#### Probability rule five

The ﬁfth basic rule of probability is also known as the general addition rule. This rule states that the probability of any one of two events happening can be calculated by taking the sum of the probability that each will happen and subtracting the probability that both will happen (Di Paola et al., 2018; Doane, 2016):

P(A or B) = P(A) + P(B) – P(A and B)

(1.11)

One should always use their best logic when solving probability problems and apply these rules in a systemic manner.

### Advanced Conditional Probabilities and Independent Events (Rules 6—7)

Now, we will cover two more advanced rules for probability, which are known as the multiplication rules for ﬁnding P(A and B). Before these rules are stated, we must ﬁrst cover two important concepts: independent events and conditional probability. Two events, A and B, are considered to be independent if the fact that one event has occur- red does not affect the probability that the other event will occur. However, when the occurence of one event affects the probability that the other event will occur, then the two events are dependent (Di Paola et al., 2018; Doane, 2016).

#### Probability rule six

The sixth basic rule of probability is also known as the multiplicative rule for inde- pendent events. This rule states that if A and B are both independent events, then the probability of both A and B occurring is the product of probability A and probability B (Di Paola et al., 2018; Doane, 2016). If A and B are two independent events, then

P(A and B) = P(A) \* P(B)

(1.12)

#### Probability rule seven

The seventh and ﬁnal basic rule of probability hinges on the fact that the probabilities of certain events may be impacted by whether or not other events have occurred. This rule is known as the conditional rule of probability. It states that the conditional proba- bility of event B, given event A, is the probability of both events A and B occurring divided by the probability of event A occurring (Di Paola et al., 2018; Doane, 2016). The conditional probability of event B, given event A, is

P(B | A) = P(A and B) / P(A)

(1.13)

### Summary

This unit exposed you to some of the most basic foundational statistical principles, including descriptive statistics, inferential statistics, and basic probability analysis. Descriptive statistics are important tools that should be a part of any data analysis, as they allow researchers to evaluate a dataset and help them to choose the most appropriate, more advanced statistical technique to incorporate into future analy- ses. Inferential statistics are a powerful set of tools that allow us to use a small sample to infer the parameters of a population. This gives us a tremendous ability to understand the world in an efﬁcient and sustainable manner, and is an impor- tant tool to study and continually use in a business environment.

There are seven basic rules of probability. One should always use their best logic when solving probability problems, and apply these rules always in a systematic manner. Sometimes, our minds seek a simple solution and can easily overlook some of the nuances of a probability problem. These are foundational principles, which could be used as the building block for more advanced studies in multivari- ate statistical tools and concepts.

# Unit 2 – Bivariate Analysis

### Study Goals

On completion of this unit, you will be able to …

… test and evaluate two means for statistical signiﬁcance.

… create and interpret cross tabulations.

… calculate and interpret correlation coefﬁcients.

# 2. Bivariate Analysis

## Case Study

The grocery store chain, Lesko, want to know if its coupons offering a ﬁve percent dis- count for select items in the store are effective at attracting incremental sales. Lesko’s chief growth ofﬁcer, Catherine Vauxhall, wonders whether the company is giving up too much of their proﬁts by offering these coupons to customers, since the store has recently seen an increase in sales. These coupons were only being offered to Lesko’s loyalty program subscribers. Catherine’s concern is that these are the wrong target cus- tomers, since they are likely to be coming into the store to shop anyway. Since all of Lesko’s current loyalty subscribers are receiving these coupons, Catherine considers the best way to set up this test. Should a “no mail” control group be created? What about the sample size? What statistical test and software should she use to ﬁnd evidence to support her hypothesis? These are the questions Catherine knows she will need to address before she receives the desired information.

This unit is an introduction to bivariate analysis, which is a series of statistical analysis procedures on two variables. This is in contrast with univariate (for a single variable) and multivariate (for three or more variables) statistics. Bivariate analysis focuses on situations with one dependent and one independent variable. The denotation of these variables are X and Y, respectively. With bivariate analyses, we analyze whether changes occur between these two variables, and to what extent.

## 2.1 Crosstabulations

**Crosstabulations**, also known as crosstabs, are essential for packaged goods ﬁrms. These ﬁrms use **market research** ﬁrms to conduct ﬁeld research and extensive surveys of consumers and current customers, using crosstabulations to segment their custom- ers and extract insights on the use of their products. One example is the use of the product in the last 30 days, divided by various age brackets. These two variables can be arranged in a table for researchers to then ascertain whether certain age groups are using the product more often (Doane, 2016).

**Market research**

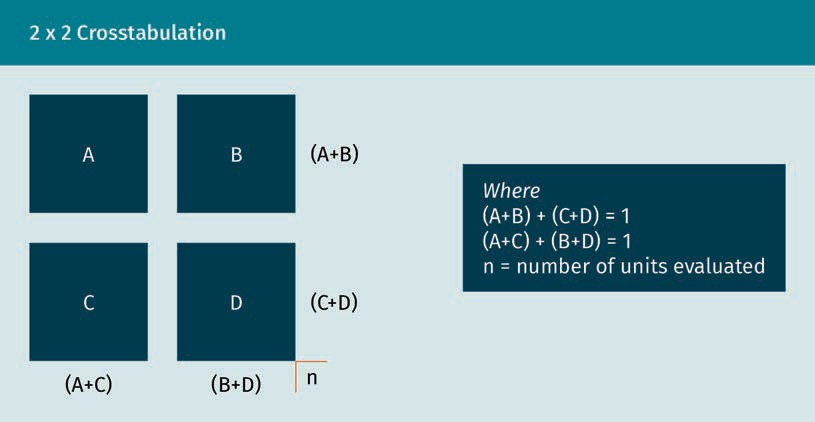
Some of the fast- moving packaged goods companies deploy market research ﬁrms to develop sophistica- ted techniques for crosstabs that can have many levels subject to sample size liminations.

**Crosstabulations**

The statistical signif- icance of a crossta- bulation is usually measured using non-paramentric statistical techni- ques, including the chi-square analysis.

A crosstab for two variables is called a two-way contingency table and forms a table with a rectangular format. The table contains rows R of the X variable, and columns C of a Y variable. Cells show each intersection, which are the possible outcomes. The fre- quency of the joint occurrences of the X, Y outcomes are contained in the cells. An R x C table refers to a contingency table having R rows and C columns (Doane, 2016).

A variable having only two categories is called a dichotomous or binary variable. When both variables are dichotomous, this results in a two by two (denoted 2 x 2) table (Doane, 2016). The ﬁgure below is an example of a 2 x 2 table.



A crosstabulation contains the three main probability distributions that we are con- cerned with: joint, marginal, and conditional (Doane, 2016). The joint probability distri- bution in a crosstab describes the proportion of the subjects jointly classiﬁed by a cat- egory of both X and Y. Dividing the cells of the crosstab by the total provides the joint distribution. The sum of the joint distribution is one, as indicated in the ﬁgure above (Doane, 2016).

The marginal probability distribution is described as the distribution of the X (row) or Y (column) variable alone. The marginal probability distributions are provided by the row and column totals of the crosstab. The sum of a marginal distribution is one, as indicated in the ﬁgure above (Doane, 2016).

The distribution of one variable, given the levels of the other variable, is described by the conditional probability distributions. The cells of the crosstab divided by the row or column totals provide the conditional probability distributions. The sum of a condi- tional distribution is one, as indicated in the ﬁgure above (Doane, 2016).

There are many other rules of crosstabulation tables, which can be further researched in the readings. These include rules about when both variables are random, when both variables have different levels of measurement, and when the Y or X variables are combinations of random and ﬁxed variables (Doane, 2016).

## 2.2 Mean Comparison Test

How do we know whether the difference between two means or medians is due to a signiﬁcant treatment effect or just random variation? We already know from the study of inferential statistics that the sample mean can vary from the population due to ran- dom sampling error. Therefore, when we see a difference between two sample means or medians, how do we know if this difference is due to random error, or some treat- ment effect or business activity that resulted in a signiﬁcant difference? Our study of mean and median comparison tests will enable us to detect this difference. A **nonparametric** test can be used, in this instance, to compare medians. This test is known as the Wilcoxon signed-rank test. Nonparametric tests are simpler to perform than parametric tests and do not rely on as many assumptions about the distribution of the data. Nonparametric literally refers to data not having a normal distribution, meaning that these tests do not depend on the data being normally distributed. The Wilcoxon signed-rank test is used with dependent samples, with these being the most interesting in the study of bivariate analysis since we are mainly focused on the rela- tionship between two variables at this stage. The Wilcoxon signed-rank test evaluates the difference between the median of two dependent samples, rather than the mean, which helps to give it the power to be used with more challenging samples because the median is less susceptible to outliers in the data (Fogarty, 2017). An example of two dependent means is the results of a pretest for knowledge in corporate ethics com- pared to a posttest taken after employees received an online ethics training. We can conclude that the two samples are dependent, or related to one another, since the same employees take the test both before and after the training.

**Nonparametric**

Statistical techni- ques that are non- parametric do not require the same assumptions of the data distributions as parametric techni- ques. However, there is a tradeoff in the power of the statisti-cal technique.

The Wilcoxon signed-rank test (also known as the Wilcoxon signed-rank sum tests) was developed by Frank Wilcoxon while doing chemical compound research at American Cyanamid Labs in Stamford, Connecticut for products, such as Old Spice, Combat Insec- ticides, and Pine Sol Cleanser (Fogarty, 2017). There are two slightly different versions of the test, but they are very similar and are both typically referred to as “Wilcoxon” tests. The basic idea of the tests is to compare the median of the two samples and see if the difference falls outside of what one would expect from the error of drawing a sample. In the example mentioned above, we would want to know whether the online ethics training had any effect on the results (Doane, 2016).

Remember, for data that are normally distributed (assuming a large enough sample), the mean and median are equal. For the purposes of the Wilcoxon test, the null **hypothesis** is that the medians of the two samples are equal. This indicates that there is no treatment effect (Doane, 2016). The research question we are testing from the example above is whether the difference between the pretest and posttest mean is statistically signiﬁcant due to the corporate ethics training.

**Hypothesis**

Hypothesis tests are critical in determin- ing whether the dif- ferences between medians are due to a treatment or pro- gram, or if they are a random samplingerror.

The prerequisites to run a Wilcoxon signed rank test are as follows:

* The pretest observation must be matched with the posttest.
* There should be no tied ranks. If there are tied ranks, a workaround will need to be executed
* The dependent variable must be continuous (other non-parametric tests are availa- ble for categorical dependent variables, e.g., the alternate test suggested below), which is the case to these being test scores (Doane, 2016).

The seven steps of the Wilcoxon signed-rank test are as follows:

1. Calculate the differences between pairs of pre- and posttest scores.
2. Rank the differences without signs.
3. Reintroduce the signs to the ranks.
4. Sum the positive and negative ranks.
5. Denote the unsigned value of the smaller sum of ranks with 6 ; n = sample size (number of ranks), + = mean, and ,2 = standard deviation.
6. Calculate *μ* = *n(n +* 1)/4, σ2 = (2*n* + 1)*μ*/6, and then *Z*= (*T* – μ)/σ.
7. Obtain the p-value (as if it were α) from the z-table for a one-tailed test as appropriate.

If the p-value is very low (i.e., less than .05), this indicates that it would be very rare for the difference between means to occur by chance alone (Doane, 2016). Hypothesis tests are, perhaps, one of the most important tools developed to aid the advancement of scientiﬁc knowledge in modern times. Many great scientiﬁc discoveries, including vac- cines and other lifesaving medicines, have been discovered via hypothesis testing.

## 2.3 Correlations

To determine the direction of the linear relationship between the two variables, the covariance can be used. The values of covariance can be any number between negative and positive inﬁnity. The value of covariance between two variables can be determined by summing the product of the differences from the means of the variables using the following equation:

(2.1)

The variances of the variables involved determine the minimum and maximum values of the covariance. The scaling of the variables and units of measurement can have a direct effect on these variances. Therefore, covariance is not very useful to determine the magnitude of the relationship.

Another statistic known as correlation is very closely related to covariance, and has some useful features, such as remaining unaffected by the change in scale, dimen- sions, and location. It can also be used to compare two pairs of variables across differ- ent domains (Doane, 2016). The correlation is determined by normalizing the cova- riance. This is accomplished by dividing it by the product of the standard deviations of the two variables (Doane, 2016):

(2.2)

Here, Cov is the covariance, σx is the standard deviation of X, and σy is the standard deviation of Y.

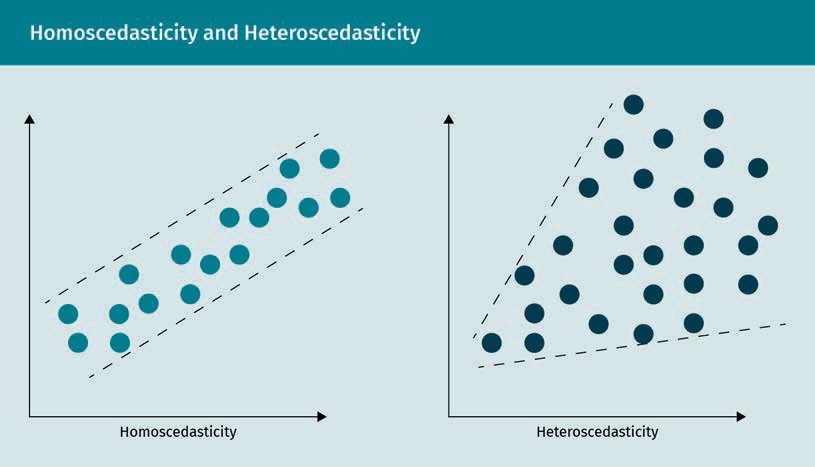
The correlation coefﬁcient measures the strength and direction of the relationship between two variables. It is a measure of the degree of the “linear” relationship between two variables. The most common correlatioin coefﬁcient is known as the Pear- son product moment correlation, noted in literature as “r” (Doane, 2016).

Correlation was originally conceived by Sir Francis Galton, a cousin to Charles Darwin, in his research on the theory of evolution. Karl Pearson, under the same research topic related to Darwin’s theory, later reﬁned the calculations leading to the Person product moment correlation. An outline of the derivation appeared in a paper Pearson pub- lished in 1896 (Fogarty, 2017).

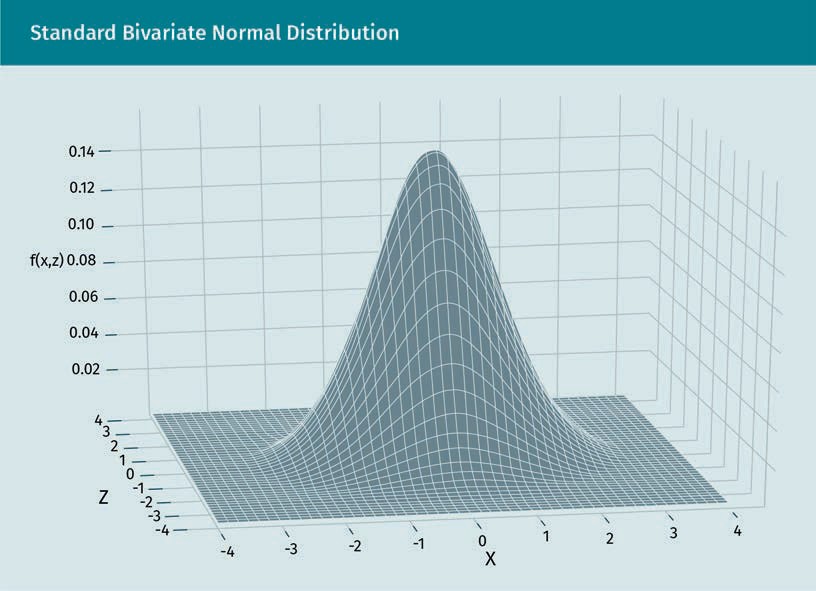
Correlation coefﬁcients range from -1 (perfectly inverse/negative correlation) to a +1 (perfectly direct/positive correlation). A correlation at (or close to) zero indicates that there is no discernable linear relationship between the variables. The squared correla- tion coefﬁcient indicates the percentage of variation in one of the variables due to knowing the other variable.

We must be aware that the correlation coefﬁcient only measures the straight-line or “linear” relationship between variables; correlation does not measure nonlinear rela- tionships very accurately. One assumption with correlation is homoscedastity. The homoscedastity assumption is that the variance is the same for all values of variables on the y-axis of a scatterplot. To assume homoscedasticity means to accept that var- iance is ﬁxed throughout a distribution. When there are serious violations of the homo- scedastity assumption, it underestimates the Pearson product moment correlation coefﬁcient (Doane, 2016).

The heteroscedasticity assumption with correlation results from a non-normality of one of the variables (as shown in the ﬁgure below), an indirect relationship between varia- bles, or a data transformation. It is important to note that heteroscedasticity does not invalidate a correlation analysis, instead, it challenges the ability of the tool to be able to measure the relationship between two variables. One can detect homoscedasticity through the use of scatter diagrams, and it is rectiﬁed through transformations of the variable (Doane, 2016).



Correlation also assumes that each pair of variables has a bivariate normal distribu- tion. In a bivariate normal distribution, each of the two variables in a pair are normally distributed, and they retain a normal distribution when added together (Doane, 2016). The bivariate normal distribution is visualized as a bell curve, as shown in the ﬁgure below.



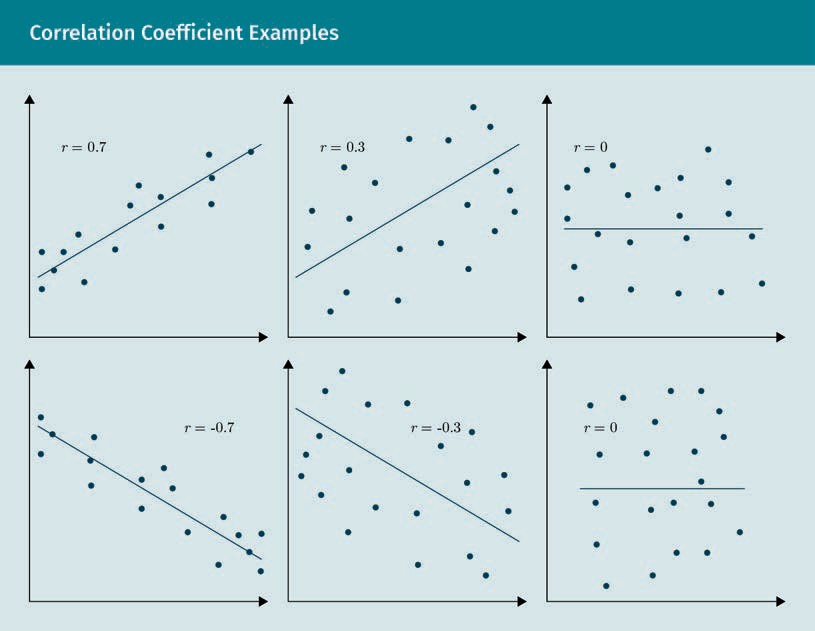
Additional issues researchers need to be aware of when dealing with correlation analy- sis are nonlinear relationships between two variables as correlation coefﬁcients are only useful for measuring linear relationships. Outliers in the data can also result in invalid correlation coefﬁcients; therefore, researchers should always check for these and take steps to either eliminate or address them prior to the correlation analysis. In additional to outliers, extreme performing groups of individuals can also affect the interpretation of correlation coefﬁcients. Sometimes, researchers combine groups in order to modify the correlation coefﬁcient, but this must be addressed carefully so as not to get the wrong interpretation. Finally, there can be truncated ranges of the varia- bles going into a correlation coefﬁcient, which could threaten the interpretation of the correlation coefﬁcient (Doane, 2016). There are three types of correlation:

1. Simple correlation between one “dependent” variable, Y, and one “explanatory” var- iable, X (Doane, 2016).
2. Multiple correlation between one “dependent” variable, Y, and multiple “explana- tory” variables (Doane, 2016).
3. Canonical correlation between multiple “dependent” variables and multiple “explanatory” variables (Doane, 2016).

Simple correlation, as previously stated, is calculated between two variables. Multiple correlation is computed between one variable and two or more variables (e.g., the rela- tionship between the body mass index of an individual as a function of height, age, and average daily caloric intake). This means that it involves the use of multiple variables (Doane, 2016).

It is important to note that correlation not only tells us not only the strength of a linear relationship (close to -1 or +1), but also the direction. In other words, a positive simple correlation indicates that increases in the X variable are associated with increases in the Y variable (and vice versa). A negative simple correlation tells us that increases in the X variable are associated with decreases in the Y variable (and vice versa) (Akoglu, 2018).

The scatterplots below are graphical representations of the different strengths and directions of linear relationships that can exist between two variables. The straight line going through the points depicts the linear relationship.



A correlation matrix is used when a researcher wants to display many different simple correlations for multiple variables. The correlations are all assembled into a table with the number one (1) always going down the main diagonal of the table. For instance, with four variables, there would be four different simple correlations computed; the matrix (table) would consist of these four correlations in the off-diagonal elements and the number one (1) in the main diagonal of the table. **Spurious correlations** can also occur when two variables seem to be correlated (numerically), but are not. Often, their correlation is driven by a third, hidden variable.

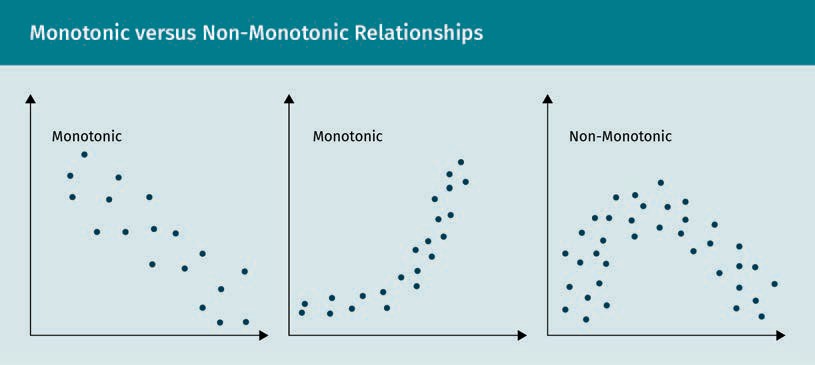
**Spurious correla- tions**

These can be pre- vented by having a strong hypothesis about why the corre- lation is occurring. Without a hypothe- sis, it is very easy to ﬁnd correlations in the data that don’t really exist.

Correlation and regression analysis go hand in hand, and one may sometimes be inter- changed for another depending on the application. While correlation measures the strength and direction of the relationship, it does not give the actual linear relation- ship. Regression analysis will yield the equation of the straight line (the blue line in the plots above) going through any scatter of points, where b0 is the “Y” intercept and b1 is the slope. The correlation coefﬁcient and the slope will always have the same mathe- matical sign (Doane, 2016).

The Pearson product moment correlation is the most efﬁcient equation when there are two numerical variables being measured. There are other types of correlations that can be efﬁcient when the two variables have mixed levels of measurement. In statistics, we usually measure three types of correlations in addition to the Pearson correlation. These are the Spearman correlation, Kendall rank correlation, and the point-biserial correlation (Akoglu, 2018).

Remember that the Pearson correlation assumes that the two variables are linearly related. The Spearman correlation is useful if you ﬁnd out that your two numeric varia- bles are not linearly related (i.e., an exponential relationship); one or both of your vari- ables are ordinal variables; or there are outliers in the data. If any of these are the case, you can still measure the strength and direction of their relationship using the Spearman rank correlation coefﬁcient, which is considered to be a non-parametric cor- relation statistic. The Spearman rank correlation coefﬁcient, which is represented by the symbol ρ, considers the ranks of the values for the two variables. It is important to note that the ρ will always be a value between –1 and 1, since Spearman’s correlation is equivalent to calculating and interpreting the Pearson product moment correlation coefﬁcient. Finally, it should be noted that Spearman’s correlation assumes that the relationship between the two variables being evaluated is monotonic. In other words, as one variable increases, the other tends to decrease or increase (but not both) (Doane, 2016). The ﬁgure below provides a visual description of monotonic versus non- monotonic relationships.



Another non-parametric version of the Pearson correlation is known as Kendall’s tau coefﬁcient. The Kendall’s tau coefﬁcient, also known as the Kendall rank coefﬁcient, uses ranks of the data to assess statistical associations. Both the Kendall’s tau and the aforementioned Spearman’s rank coefﬁcient are good alternatives to the Pearson cor- relation when one or more assumptions of the Pearson correlation can not be met. Kendall’s tau is often used as an alternative to the Spearman correlation when the data have many tied ranks or there is a small sample size. In the latter case, p-values with Kendall’s tau are more accurate (Akoglu, 2018).

While other types of correlation coefﬁcients (including the Pearson product moment correlation) use the observations as the basis of the correlation, Kendall’s correlation coefﬁcient uses pairs of observations and assesses whether there is a pattern of con- cordance or discordance between the pairs, determining the strength of association based on the pattern observed (Akoglu, 2018). The term concordant means that pairs are ordered in the same way (consistent). A pair of observations is considered concord- ant if (x2 — x1) and (y2 — y1) have the same sign. Alternately, discordant pairs are ordered differently (inconsistently). A pair of observations is considered discordant if (x2 — x1) and (y2 — y1) have opposite signs.

The assumptions associated with using Kendall’s rank correlation are that the variables are measured on an ordinal or continuous scale and that there is a monotonic rela- tionship between variables (Akoglu, 2018). Correlation analysis in its many forms is an important component to bivariate statistical analysis and is included in many research studies. Moreover, as humans we are naturally forming associations in our own minds based on our experiences. Correlation is a way to mathematically quantify these “hunches” we are creating on a continuous basis in our professional and personal lives.

### Summary

In this unit, bivariate statistics were covered, which gave a ﬁrst look at how to measure and describe the relationship between two variables. The use of crossta- bulations is one of the most simple ways to quantify a relationship between two or more discrete variables. We looked at correlations between categorical variables; the differences between means; and the correlation coefﬁcient, which measures the direction and strength of the relationship between two numeric variables.

The Wilcoxon signed-rank test was introduced to help us compare medians. The Pearson product moment correlation measures the strength and direction of the relationship between two variables. In situations where the Pearson correlation is not applicable (i.e., with more than two variables), we may need to use the Spear- man correlation, Kendall rank correlation, or the point-biserial correlation. In the modern world of big data, bivariate statistics quantifying relationships between variables are needed to understand factors that may be related to phenomena we observe in business and our personal lives.

# Unit 3 –Probability Distributions?

### Study Goals

On completion of this unit, you will be able to …

… identify and understand discrete and numeric random variables.

… identify the properties of the standard and general normal distribution.

… interpret the student’s t-distribution.

# 3. Probability Distributions

## Case Study

The biotechnology ﬁrm, Riedling, specializes in creating better healthcare solutions for an aging population. The company has recently been gathering data around its solu- tions to better assess the long-term efﬁcacy of its treatments. Previously, the company had gathered a lot of this information from its clinical trials. However, now that the treatments are on the market, it wants to gain a longer-term view.

The data gathered encompass all of the physical information from the patients of the trial, including height, weight, age, gender, and marital status, among others. Health measurements, such as blood pressure, cholesterol, and blood sugar levels were also added. Data were also gathered on the health outcomes of the individuals, including the onset of certain critical illnesses, including diabetes, cardiopulmonary conditions, and many other clinical conditions.

Harvey Kramer, the ﬁrm’s chief biostatistician, was responsible for this initiative. He wanted to run some sophisticated statistical analyses on the data; set out the ﬁrst task of classifying all of the variables as discrete or continuous; and calculate the means, standard deviations, and other descriptive statistics. He had learned this from his training in descriptive statistics. However, this time, he wanted to go a step further and produce histograms for all of the data, knowing that if his data were statistically nor- mal, he could use techniques like correlation analysis and general linear models. He immediately noticed that variables like height, weight, and blood pressure were nor- mally distributed, while other measurements, such as health engagement, were not normal. Harvey began to create a new spreadsheet of all the variables that were nor- mally distributed versus those that were not as the next step, before beginning the deeper analysis.

## 3.1 Random Variables and their Distributions

There are many random phenomena in the world we live in. Some scientists who were trained under classical mechanics would argue that there is no true randomness in the universe, and that “random” as we know it is just the lack of information about the uni- verse. However, scientists trained with the quantum physics paradigm would argue that there is true randomness in the universe, and even have quantum experiments to lend evidence to this conjecture. Whatever the orientation toward this discussion, we as humans will experience much randomness in our daily personal and professional lives.

In statistical terms, randomness is recorded and measured in the form of random vari- ables. Typically, **random variables** are those derived from experiments; however, they can also be derived from processes happening on a continuous basis around us. Ran- dom variables are critical components to statistical analysis and knowledge of them is absolutely required to gain a full understanding of probability distributions (McEvoy, 2018).

**Random variables**

Unlike algebraic vari- ables, which have a ﬁxed value, random variables have an entire set of values and could possibly take on any one of these values.

Random variables make it easy for us to quantify random processes and perform stat- istical calculations. One of the most typical random variables is the coin toss. If we toss the coin one hundred times and record each outcome in a journal, the recorded out- come is a random variable. Random variables are classiﬁed as one of two types: dis- crete or continuous. Discrete random variables, such as the number of students in a classroom, can be determined by counting. Continuous random variables, on the other hand, can only be determined through some type of measurement. An example of a continuous random variable is time, which can be broken down, for example, into cen- turies, decades, years, days, hours, minutes, seconds, and miliseconds (McEvoy, 2018).

Random variables emerging from experiments, which can be either continuous or dis- crete, are important in statistics because these variables all follow distributions that have properties that useful when conducting statistical analysis. Examples of discrete random variables include the number of cats in a litter, people attending a basketball game, students in attendance in class, and people in line at a department store on a given time and day. Continuous random variables, on the other hand, cannot be coun- ted and can take inﬁnitely many values. In other words, continuous random variables have inﬁnite precision and can only be determined through some type of measure- ment. Examples include distance, temperature, pressure, height, weight, mass, density, temperature, and volume.

It is useful to note that there are no absolutes. Discrete random variables can be parti- ally continuous and, moreover, continuous random variables can be partly discrete. Such variables are referred to as mixed type variables.

## 3.2 Discrete Probability Distributions

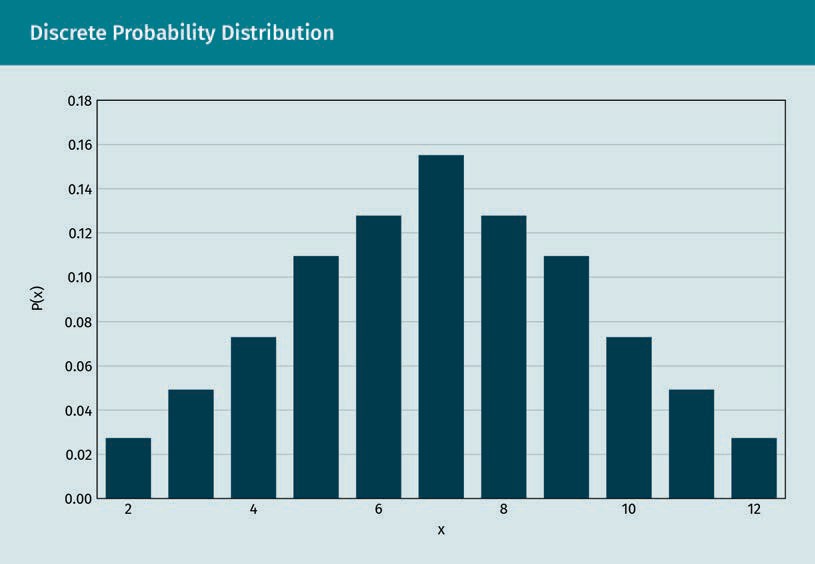
Discrete random variables have a distribution that is identiﬁed as a **discrete probabil- ity distribution** (Viti et al., 2015; McEvoy, 2018). The discrete probability distribution of a discrete random variable is a catalog of probabilities associated with each of its possi- ble values. It is also sometimes called the probability mass function or probability function. Some of the most popular discrete probability distributions include the bino- mial probability distribution, hypergeometric probability distribution, multinomial probability distribution, negative binomial distribution, and Poisson probability distri- bution (Viti et al., 2015; McEvoy, 2018).

**Discrete probability distribution**

A discrete probabil- ity distribution is the probability of the occurrence of each value of a discrete random variable, and each possible value of the discrete random variable can be associated with a probability that is greater than zero.

In a discrete probability distribution, each possible value the random variable can assume is listed in conjunction with its probability. The following conditions must be satisﬁed in order for a probability distribution to be valid. Let x be a discrete random variable with possible outcomes x1, x2, …, xn. Firstly, each value of the discrete random variable has a probability between 0 and 1, inclusive 0 ≤ . x ≤ 1, 2. Moreover, the sum of all the probabilities is ∑. x = 1 (Viti et al., 2015; McEvoy, 2018).

Discrete probability distributions can be visualized using a histogram, such as the his- togram below, depicting the distribution of a discrete random variable. This histogram is obtained by letting X be a discrete random variable that has more than one possible outcome. We will ﬁrst plot the probability on the y-axis and the outcome of the discrete variable on the x-axis. If we repeat the experiment over many trials, and record and plot the probability of each possible outcome every time, we get a plot that represents the probabilities. This plot is called the probability distribution (PD). The height of each bar in the graph for X gives the probability of that particular outcome (Viti et al., 2015; McEvoy, 2018).



We have now learned that a discrete probability distribution describes the distribution of a discrete random variable. A discrete random variable is a variable resulting from an experiment with discrete number or countable values. For example, counting arrivals of aircraft at an airport is a discrete random variable. Each value of a discrete random variable has a probability which is between 0 and 1, inclusive. We have now learned that a discrete probability distribution describes the distribution of a discrete random variable. A discrete random variable is a variable resulting from an experiment with discrete number or countable values. For example, the number of arrivals of aircraft at an airport is a discrete random variable. Each value of a discrete random variable has a probability between zero and one, inclusive. Moreover, all of the probabilities will sum to Shape

Description automatically generated with medium confidence.

Discrete probability distributions can be visualized via a histogram or probability distri- bution plot. This histogram is obtained by recording the probability of each possible outcome of a discrete random variable over many trials, then plotting the probability on the y-axis and the outcome of the discrete variable on the x-axis. There are many types of discrete random variables, including the binomial probability distribution and the hypergeometric probability distribution (Viti et al., 2015; McEvoy, 2018).

3.3 Continuous Probability Distributions

If the random variable is continuous, then its distribution is classiﬁed as a **continuous probability distribution**. The probability that a continuous random variable will assume any particular value in a distribution is zero, which is a key differentiator between a continuous and a discrete probability distribution (Crooks, 2019). As a consequence, a continuous probability distribution must be described using an equation or formula, rather than in tabular form like a discrete probability distribution (Crooks, 2019).

**Continuous proba- bility distribution**

These are typically plotted as a curve, the area of which is the cumulative prob- ability. They typically represent the proba- bility by utilizing speciﬁc formulas, geometry, technol- ogy, or probability tables.

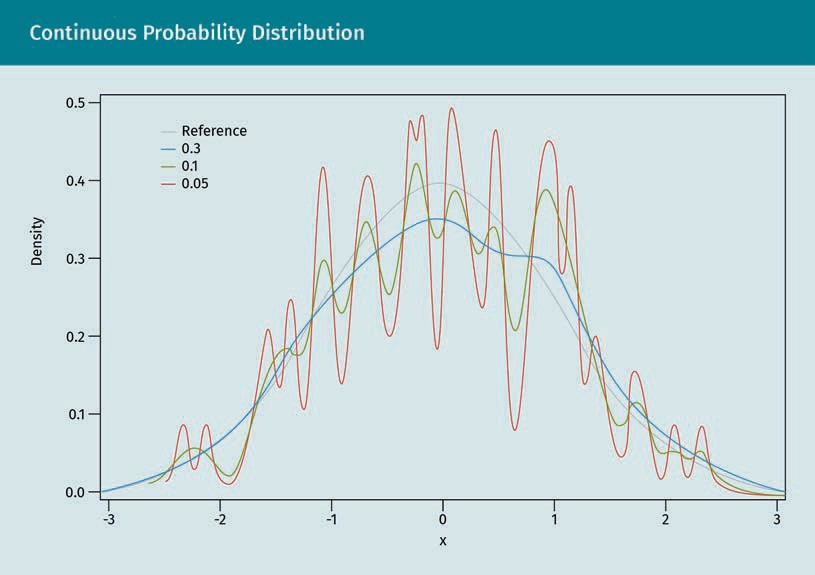
A discrete probability distribution has a range of countable values. For example, if a card company wanted to create birthday cards for all possible ages, they would range from zero to 122, with 122 representing the oldest person that ever lived. A continuous distribution of random variables, on the other hand, has a range of values that are inﬁnite, and therefore not countable. A continuous distribution of a random variable can also assume all possible values in the possible range of the random variable.

For example, suppose the temperature in New York City in the month of July in the past ten years has always been between 34 and 44 degrees centigrade (°C). The temperature can take any value between the ranges 34 and 44 degrees centigrade. The temperature on any day may be 39.15°C or 40.15°C, or it may take any value between 39.15°C and 40.15°C. When we say that the temperature is 39°C, it means that the temperature lies somewhere between 38.5°C and 39.5°C. Any observation which is taken falls in the inter- val (Crooks, 2019).

In addition to the differences between discrete and continuous probability distribu- tions, a formula is used to express a continuous probability distribution (Crooks, 2019). While discrete probability distributions are often described with frequency graphs or tables, continuous probability solutions are expressed using a formula known as the **probability density function** (Crooks, 2019). The ﬁgure below is an example of a typical continuous probability distribution with the continuous random variable on the x-axis and the probability density function on the y-axis.

**Probability density function**

Since many of these distributions are not closed-form solu- tions, integral calcu- lus is used to ﬁnd the area under the curve, as represen- ted by the probabil- ity density function.



In terms of types of continuous probability distributions, the normal distribution remains the the most common in the category. In addition to representing the distribu- tion of many continuous random variables in biology social sciences and business, it can also approximate the Poisson distribution, as well as the binomial and hypergeo- metric distributions (Shakil et al., 2010). Additional continuous distributions, which will not be covered in this section but are, nevertheless, important for speciﬁc applications include the beta, Cauchy, exponential, gamma, logistic, and Weibull distributions. There is also a slew of less common continuous distributions in the literature, for example, the Shakil-Singh-Kibria distribution, based on the Whittaker functions (Shakil et al., 2010).

The probability that a continuous random variable will assume a particular value is zero, thus representing a key differentiator between a discrete and continuous proba- bility distribution. Consequently, a continuous probability distribution is typically repre- sented using an equation or formula rather than using tables as with a discrete proba- bility distribution (Crooks, 2019). This formula is often referred to as the probability density function. Continuous random variables are some of the most important meas- urements of human and business activity on the planet, and continuous probability distributions are critical to the understanding and prediction of these variables. How- ever, it is important to remember that, given the nature of continuous probability dis- tributions, working with discrete distributions is much easier, which motivates many researchers to begin their analysis with a discretized version of a continuous variable (Crooks, 2019).

## 3.4 Normal Distribution

In early agricultural studies, it was discovered that if plants were continuously bred with superior varieties, they wouldn’t continue to improve indeﬁnitely. Instead, they would vary and create an average variety that would have outliers, but most breeds would hover around the center of distribution. This was a principle known as regres- sion toward the mean, and was ﬁrst discussed by Francis Galton, who ﬁrst noticed this in agricultural experiments before extending the study to the heights of parents and their children. P. L. Chebyshev was a Russian mathematician of the nineteenth century who ascertained that the percentage of observations falling between two distinct val- ues, whose differences from the mean have the same absolute value, is related to the variance of the population. He then developed Chebyshev’s theorem, which provides a conservative estimate of the above percentage (Fogarty, 2017). For any population or sample, at least

(3.1)

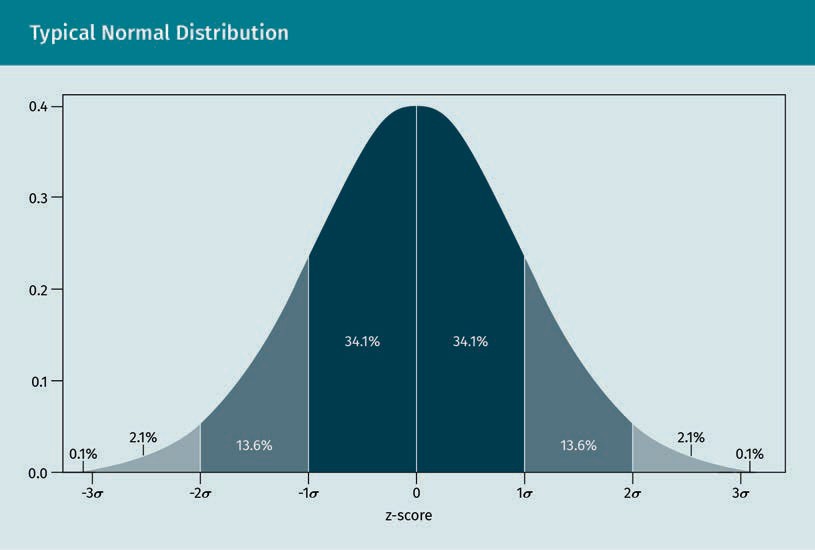
of the observations in the dataset fall within k standard deviations of the mean, where k ≥ 1.

Finally, the mathematicians Gauss and Adrian independently derived the formulas for the **normal distribution**. However, Gauss got the credit for this, and another name for the normal distribution is the Gaussian distribution (Fogarty, 2017). Since the normal distribution is not a closed-ended solution, one needs to use integral calculus to ﬁnd the true probabilities. Therefore, the hypothetical normal distribution probability tables are used to ease the calculations. These probabilities represent the area under the normal distribution, with one equaling the entire area and all of the individual values, depending on where the value lies on the distribution in relation to the mean (Fogarty, 2017).

**Normal distribution**

Population parame- ters deﬁne the shape and probabili- ties of any probabil- ity distribution, including the normal distribution. The shape of the normal distribution is deﬁned by two parameters: the mean and standard deviation.

The normal distribution is bell-shaped and symmetric around the mean. This means that the left side of the center is an identical image of the right side. The standard deviation determines the thickness of the distribution, with higher standard deviations resulting in a much skinnier curve. The values of the mean and standard deviation parameters in the normal distribution change the shape of the distribution, resulting in multiple forms (Attenborough, 2003). Below is an example of a typical normal distribu- tion.



In terms of importance, the normal distribution is perhaps number one in the study of statistics because many continuous random variables in business and everyday life display this bell-shaped curve when compiled and graphed. This is the reason that the normal distribution is often referred to as the bell curve (Doane, 2016).

Parametric statistics require data to be normally distributed. These are among the most powerful techniques used in business and the social sciences. If the data are not nor- mally distributed, then researchers will have to either transform the data or have to use a less powerful type of statistical test, known as non-parametric statistics (Doane, 2016; McEvoy, 2018).

The majority of the continuous data values in a normal distribution will tend to cluster around the mean. The further the observation deviates from the mean, the lower the likelihood of an event occurring. The tails of the normal distribution move out into inﬁnity and are asymptotic, which means that they approach, but never quite meet, the x-axis, or horizon (Doane, 2016).

Since the deviations from the normal curve are predictable in terms of the probability of deviating from the mean value, the area under the normal distribution represents the probability, and the total area under the normal curve represents all of the proba- bilities, therefore summing to one (Doane, 2016). Interestingly, in a perfect normal dis- tribution, the mean, median, and mode will be the same, or very similar, values and lie at the very peak of the curve. There are speciﬁc tests for normality of variables that take this principle into account by calculating and comparing the mean, median, and mode (Doane, 2016).

Oftentimes, students will come across the term “standard normal distribution,” but what is the difference between a normal distribution and a standard normal distribu- tion? A normal distribution is determined by two population parameters: the mean and the variance. These can take on any values, but share the same probabilities regardless. Taking this into account, a normal distribution can be standardized to have a mean of zero and a standard deviation of one. This standardization creates what is referred to as the standard normal distribution (Doane, 2016). The values, or raw scores, are typi- cally standardized by turning them into z-scores. This standardization procedure ena- bles researchers to ascertain the proportion of value that falls within a speciﬁed stand- ard deviation from the mean. This way, they are able to use the empirical rule (Doane, 2016; McEvoy, 2018).

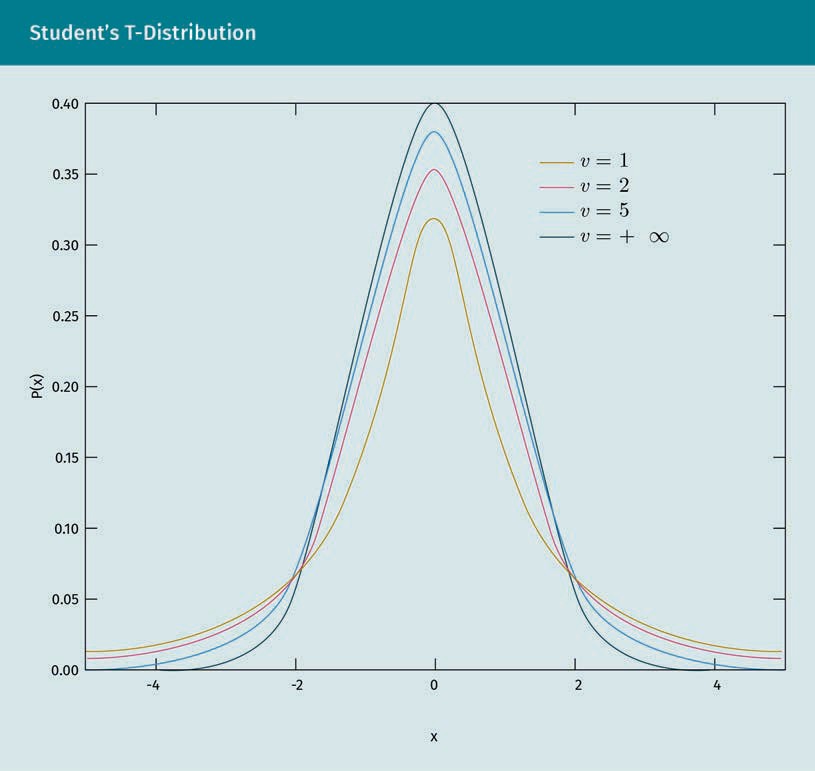
The empirical rule is a unique property of the normal distribution. It’s actually also referred to as the “95% rule” since this is the most frequently used interval in hypothe- sis tests. The 95% rule states that 95 percent (actually 95.44 percent) of observations fall within two standard deviations of the mean on a normal distribution (Doane, 2016; McEvoy, 2018).

The empirical rule expands this deﬁnition by stating that, in a normal distribution, around 68 percent of data will be within one standard deviation of the mean, around 95 percent will be within two standard deviations of the mean, and around 99.7 percent will be within three standard deviations of the mean. Therefore, in statistics, we often use the 68–95–99.7 rule as shorthand to remember the percentages of observations that fall into the three standard deviations from the mean in a normal distribution (Doane, 2016).

3.5 T-Distribution

The popular Pearson product moment correlation was ﬁrst developed and published by Karl Pearson, the British biostatistician and mathematician who was a protégé of Fran- cis Galton and developed statistical techniques to validate Darwin’s Theory of Evolu- tion. One of Pearson’s own protégés was William Sealy Gossett, who was a gifted statis- tician, hired by the famous Guinness Brewing Company to improve the quality processes at the brewery. Gossett found that the present statistical techniques that assumed large samples did not work very well in the applications he was working on at Guinness, which had a very small sample size. He therefore engaged in research around testing a statistical hypothesis with small samples, and developed knowledge around the t-distribution (Fogarty, 2017). Gossett corresponded with Karl Pearson, most famous for developing the Pearson product moment correlation.

Karl Pearson saw the value of the work that Gossett was doing to the broader scholarly community, as opposed it to being proprietary only to Guinness Brewery. The communi- cation of this type of information to scholars is primarily done through scholarly jour- nals or at academic conferences. Since Gossett did not want Guinness Brewery to know that he was publishing some of the statistical work he was doing at the brewery, he published his work under the pseudonym “Student.” Hence, we now know why the t- test is referred to as Student’s t-test. We are very lucky that Gossett published his work under a pseudonym despite facing the risk of becoming employed, as hypothesis test- ing, which includes the use of t-tests is regarded as one of the most important discov- eries of the nineteenth century (Fogarty, 2017). The illustration below depicts a typical t- distribution.



When compared with the shape of the standard normal distribution one can see that the t-distribution is still bell-shaped but has heavier tails. The t-distribution is deﬁned by the degrees of freedom. This is a correction related to sample size (Doane, 2016; Tat- tar, Ramaiah, & Manjunath, 2016; McEvoy, 2018).

### Degrees of Freedom

In applied statistics, the df, or **degrees of freedom**, is the number of independent val- ues that can vary in your analysis without violating any constraints. This concept is an important one that is used in many techniques, including the t-distribution and t-tests, and can impact the precision and power of one’s analyses (Fogarty, 2017; Frost, 2019).

**Degrees of freedom**

This concept was ﬁrst mentioned by Carl Frederick Gauss in the 1880s, and was brought into modern statistics by William Sealy Gos- sett in the early1900s

Degrees of freedom (typically a positive whole number) are usually calculated by taking the sample size and subtracting the number of parameters needed (Frost, 2019).

The idea of degrees of freedom is that the the quantity of independent information you have constrains the number of parameters that you can estimate (Frost, 2019). Using this logic researchers always desire more information to go into parameter esti- mates and have more powerful statistical analyses. Therefore, the more degrees of freedom, the better (Frost, 2019).

### Comparison of the T-Distribution with the Normal Distribution

Researchers are limited in their ability to choose the normal distribution (which requires the population standard deviation) and will default to using a t-distribution (which only requires using the sample standard deviation) for their hypothesis testing (Virginia Tech, 1999). However, since t-tests are recommended with sample sizes of less than 30 observations, researchers must take sample sizes into account when thinking about their critical values. However, the t-distribution exhibits asymptotic properties and will exhibit almost the same critical values as the normal distribution when sam- ple sizes exceed 30 observations (Virginia Tech, 1999).

However, the t-distribution is still very similar to the normal distribution in that it has a precise mathematical deﬁnition. Other similar features include the smooth shape of both the normal distribution and the t-distribution, and that the normal distribution and the t-distribution are both symmetrical (each side of the mean will mirror the other). Both distributions also have a mean of zero. The ultimate similarity is that, as the sample size approaches inﬁnity (i.e., becomes larger), the t-distribution very closely resembles the normal distribution. (Doane, 2016; Tattar et al., 2016; McEvoy, 2018).

## Summary

Random variables make it easy for us to quantify random processes and perform statistical calculations. Random variables emerging from experiments, which can be either continuous or discrete, are important in statistics, because these variables all follow distributions that have properties that are useful when conducting statis- tical analysis. Discrete probability distribution describes the distribution of a dis- crete random variable, which is a variable resulting from an experiment with dis- crete number or countable values. Discrete probability distributions can be depicted using a histogram or probability distribution plot. If the random variable is continuous, then its distribution is classiﬁed as a continuous probability distri- bution.

The normal distribution recorded by Gauss is bell-shaped with probabilities repre- senting the area under the curve; the entire area and all of the individual values is equal to one, depending on where the value lies on the distribution in relation to the mean. The empirical rule states that, in a normal distribution, around 68 per- cent of data will be within one standard deviation of the mean, around 95 percent will be within two standard deviations, and around 99.7 percent will be within three standard deviations.

A t-distribution (i.e., Student’s t-distribution) describes the standardized distances of sample means (in terms of standard deviations) to the population mean when the population standard deviation is not known (and only the sample standard deviation is available for use), and the observations come from a normally distrib- uted population.

# Unit 4 – Statistical Estimation Methods

## Study Goals

On completion of this unit, you will be able to …

… compute and analyze a point estimate.

… compute and analyze an interval estimate.

… decide when to use a conﬁdence interval versus a single point estimate.

# 4. Statistical Estimation Methods

## Case Study

The Swiss company, Accurate Timepieces, is a maker of luxury watches, and had recently done an initial public offering (IPO) and had shares traded on the Swiss stock exchange (SIX). The chief executive ofﬁcer (CEO), Hermann Marx, began working with Marie Schultz, the chief ﬁnancial ofﬁcer (CFO), to create an annual shareholder report. Hermann and Marie calculated many key performance indicators (KPIs) for the organi- zation from the business operations and ﬁnance, and wanted to report some of these ﬁgures in the annual report.

However, Hermann noticed from the analysis that quite a few of these measures con- tained a high degree of variability and, therefore, a single point estimate (i.e., average) may not be suitable for reporting. Doing this would be somewhat deceptive to share- holders. Marie had a great suggestion that any time some of the KPIs displayed a high degree of variance, as measured by their standard deviation, they would create conﬁ- dence intervals and report a range of values, within which the population parameter could reside. Hermann thought this was a great idea. Now all they needed to do was decide on the conﬁdence level for the calculations. “This is an easy decision,” exclaimed Hermann. “We are a Swiss watch company and precision matters. Therefore, it must be at least a 95 percent conﬁdence level,” Marie said, agreeing.

4.1 Point Estimation

**Point estimators** are metrics or functions that are utilized in inferential statistics to ascertain an approximate value of a population parameter from a random sample selected from a population. Point estimators extract the data from a sample of the population to calculate a point estimate or a statistic that serves as the most likely estimate of an unknown population parameter. Point estimators can be a variety of sta- tistics, including means, proportions, medians, and standard deviations (Doane, 2016; McEvoy, 2018).

**Point estimators**

A single value or point is used to draw inferences about a population by esti- mating the value of an unknown param- eter in a point esti-mator.

A point estimate can be useful if a researcher wants a single number to be representa- tive of the entire population. This is only possible where central tendency is present, such as in a normal distribution where the most likely value will be the mean. The probability of assuming any particular value is zero in a continuous probability distri- bution, so researchers attempt to generate a value that is as close as possible to the population parameter. If this objective is met, then the point estimator is deemed an unbiased estimator. Other characteristics used to evaluate an estimator include consis- tency and efﬁciency. Managers measuring performance favor point estimates because they can assign people, processes, and businesses a goal based on the measure. How- ever, managers are cautioned that the results may just be due to random ﬂuctuation without a hypothesis test or using an interval estimate (Doane, 2016; McEvoy, 2018).

There are diverse types of point estimators, each of which contains different properties in relation to being reliable estimators of the population. Some of the more popular ones include the Bayes estimator, minimum-variance mean-unbiased estimator (MVUE), best linear unbiased estimator (BLUE), minimum mean squared error (MMSE), maximum likelihood estimator (MLE), median-unbiased estimator, method of moments, and generalized method of moments (Lehmann & Casella, 1998).

A point estimate is used when an analyst desires a single point of the most likely value of a population parameter. This can be any univariate statistic, including the mean or even a proportion. However, it should be noted that the point estimate calculated from a sample of the population will always vary based on random sampling error. It is always recommended that a point estimate is presented along with some measure of dispersion (i.e., standard deviation) so that the consumer of the statistic will have a gauge as to how reliable that estimate is, given the amount of variation in the popula- tion or sample.

Take, for instance, a human resource manager who is asked by a candidate about the average salary of the ﬁrm. If the CEO and CFO of the ﬁrm, who are the founders, both make $500,000 per year, and a random sample of the 20 other employees of the com- pany who make less than $50,000 per year is calculated, then the average salary of the ﬁrm (point estimate) would be around $175,000. This is a deceiving ﬁgure. Now, the human resource ofﬁcer can present the median salary, which is not subjected to outli- ers, or they can present the mean using an interval estimate.

4.2 Interval Estimation

An **interval estimate**, as opposed to point estimate, can be useful if a researcher decides that a point estimate is not representative as a population inference due to a skewed distribution or from being uncomfortable (from a risk perspective) reporting a point estimate. Examples from the latter include business managers who fear they may be judged on their performance when making a point estimate, when a range of esti- mates would be more appropriate from a statistical perspective.

**Interval estimate**

An interval estimator draws inferences from a sample about a population by esti- mating the value of an unknown popula- tion parameter using an interval rather than a single point. With an interval esti- mator, we attempt to construct an interval that contains within it the actual popula- tion parameter with a speciﬁed probabil- ity

In this case, the interval estimate would be used. The interval estimate is a range of values, within which the actual population parameter has an estimated probability of falling. A conﬁdence interval is one of the most popular interval estimation techniques. Conﬁdence intervals can be calculated for a variety of statistics, including means and proportions. Interestingly, conﬁdence intervals can even be used to conduct hypothesis testing. In addition to being a range of values rather than a single point estimate, inter- val estimators are also different from point estimators in that they reﬂect the effects of sample sizes, with larger sample sizes tightening the width of the interval by reducing the standard error. For example, suppose we are trying to estimate the income of work- ers in a factory. An interval estimate would state that the (unknown population) mean income per hour is between $31 and $48 with a probability of 0.95. (Greenﬁeld et al., 1998; Altman & Bland, 2011). The formula for conﬁdence intervals (CIs) for the mean is depicted below:

(4.1)

where

CI = confidence interval

{"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\"><menclose notation=\"top\"><mi>x</mi></menclose></math>"}
top enclose x = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

Oftentimes, the characteristic being measured is not continuous, but is a categorical variable. Examples of this include political afﬁliation, gender, and marital status. In this instance, the objective would be to estimate a population proportion, *p*, using a sample proportion, , plus or minus a predetermined margin of error. This activity will result in a conﬁdence interval for the population proportion, *p* (Greenﬁeld et al., 1998; Altman & Bland, 2011). The formula for the CI for the proportions is

(4.2)

where:

= sample proportion

z\* = confidence level value from the standard normal distribution table

n = sample size

Once you calculate the z-score, it can be looked up in a z-table, which is partially depic- ted below. For access to the full table, see the literature by Farber and Larson (2017).

|  |  |
| --- | --- |
| z\*–Values for Various Conﬁdence Levels | |
| Conﬁdence level | z\*-value |
| 50% | 0 |
| 70% | .5 |
| 80% | 1.28 |
| 90% | 1.645 (by convention) |
| 95% | 1.96 |
| 98% | 2.33 |
| 99% | 2.58 |

## Summary

A point estimate is used when searching for a single point with the most likely value of a population parameter. This can be any univariate statistic, including the mean or a proportion. However, it should be noted that the point estimate calcula- ted from a sample of the population will always vary based on random sampling error. This variation is taken into account when calculating an interval estimate, as the interval estimate (i.e., conﬁdence interval) will return a range of values, into which the population parameter is likely to fall, taking into account random sam- pling error.

It is always recommended that a point estimate is presented along with some measure of dispersion (i.e., standard deviation) so that the consumer of the statis- tic will have a gauge as to how reliable that estimate is, given the amount of varia- tion in the population or sample. An interval estimate can address this all in a sin- gle statistic, although it’s not really a single statistic, but a range of values. That said, it is certainly easier to interpret than a point estimate.

## 5. Hypothesis Testing

## **Study Goals**

On completion of this unit, you will be able to …

… identify and understand the need for hypothesis testing.

… conduct and construct a hypothesis test.

… interpret a hypothesis test.

## Case Study

The auto manufacturer the Black Forest Motor (BFM) is anticipating the implementation of a new onboard computer interface system in response to customer complaints. The new system is purported to be more user friendly when drivers want to obtain informa- tion from the vehicle or interact with and adjust certain systems, including climate con- trol. The challenge is that the cost of modifying the system across all vehicles is sub- stantial.

Hannah Brekker, the chief engineer of BFM, is discussing with Gerald Meier, the chief statistician, about how to conduct a test without making a big investment in retooling the plants. The decision is to custom engineer a small subset of cars and then provide these new cars to a random sample of customers who ordered the vehicle. BFM will then monitor the customer satisfaction scores over the next 24 months to see if the new onboard computer interface yields signiﬁcantly better results. In order to deter- mine statistical signiﬁcance, Gerald Meir recommends a t-test be performed on the data. However, he needs to discuss a few things with Hannah Brekker, including the sample size; level of signiﬁcance; 24-month observation window; and logic of using cus- tomer satisfaction scores over other options, such as net promoter scores. Hannah Brekker feels a little overwhelmed learning about all these new concepts, but, at the same time, is glad that this new interface can be successfully tested without making big investments. She reﬂects on how important hypothesis testing is as a tool available to organizations like BFM.

## 5.1 Hypothesis Testing when Population Standard Deviation is Known

Whether the population standard deviation is known or unknown makes a big differ- ence when hypothesis testing. This knowledge determines the probability distribution, which is used to conduct the test. Practically speaking, we almost never know the pop- ulation standard deviation, but it makes sense to study this since it facilitates a better understanding of the concept of hypothesis testing. Therefore, we will address the his- tory, overview, and some examples of hypothesis testing.

## Origins of Hypothesis Testing

Salsburg (2001) describes a tea party at the University of Cambridge, England, at which a lady made a claim that the taste of a cup of tea poured into milk can be distin- guished from milk poured into tea. Most of the members of the party disagreed with the lady on the grounds that her conjecture lacked a scientiﬁc basis, but Ronald Fisher, another party attendee who would later become knighted and known as Sir Ronald Fisher, proposed a type of test that could satisfy both the lady and her challengers. His test included subsequent randomized presentations of cups of tea to the lady in which the milk was introduced ﬁrst and vice versa, while recording the lady’s responses for each cup. This test he proposed was a hypothesis test, and the work of Fisher and his contemporaries, including Karl Pearson and William Sealy Gossett, created a testing strategy that has been responsible for helping to usher in some of the greatest discov- eries known to humankind (Salsburg, 2001).

So why do we need this type of testing? Well, one application is the simple security that when we make a conclusion about a phenomenon, it is not occurring by chance, and is instead related to the variable in question. In the case of the tea, the variable is the timing of adding the milk. We needed such a test in this case because the lady could have simply been lucky in guessing which tea was which. Fisher’s proposed test, however, was able to measure whether the lady’s chances of getting the number of cor- rect responses was beyond normal chance occurrence (Salsburg, 2001). The key to hypothesis testing is understanding that associations between two variables may appear to be correlated, but, in reality, are sometimes the result of randomness. In the following section, you will be exposed to the various types of hypothesis tests used in business and the social sciences today. This will allow you to test your own hypotheses about the world around you to see if your conjectures are real or simply due to random variation.

## Generalized Approach to Statistical Hypothesis Testing

Statistical hypothesis testing is a form of statistical inference in which conclusions about a population are drawn on the basis of a sample obtained from that population. Hypothesis testing provides a convenient and straightforward framework and method- ology for ascertaining the reliability or strength of evidence based on insights extracted from the sample. These results are then considered in relation to the population from which the sample was drawn. However, since it is just a sample of the population, any ﬁnding derived from that sample could be biased due to random sampling error. Hypothesis tests closely follow the scientiﬁc method, in which the researcher formu- lates a speciﬁc hypothesis based on an assumption or educated guess. The researcher then draws a sample from a population, evaluates data from the sample, and uses these data to decide whether they support the speciﬁc hypothesis.

## Hypothesis Testing Fundamentals

Hypothesis testing always starts with a research question or hypothesis, which is con- verted into a hypothesis test. This research question should ideally be based on an educated guess or theory in order to prevent a false positive. Hypothesis testing is about creating and testing two hypothesis statements about a probability distribution based on data observed from a sample distribution. A step-by-step process is deployed, enabling inferences about a population parameter using a sample and tak- ing into account the expected results if the underlying research hypothesis were cor- rect (Friedman, 2018). The ﬁve core steps to conducting a hypothesis test are as follows:

1. State the null hypothesis, i.e., H0 = 0.
2. Choose the appropriate distribution.
3. Ascertain the rejection and non-rejection regions.
4. Compute the value of the test statistic.
5. Make a decision based on the data.

**Step 1: State the null hypothesis**

In this step, you will create two hypothesis statements to determine the validity of a statistical claim; these statements will contain a null hypothesis and an alternative hypothesis. The null hypothesis makes a statement claiming a zero difference, i.e., a null. The null hypothesis undergoes the testing procedure, regardless of whether it is the original claim. H0 is the statistical notation for the null hypothesis. This represents what is assumed to be true, that which is the status quo, and always contains the equal sign. H0 = 0 is the statistical language used to represent the null hypothesis (Friedman, 2018).

It now stands to reason that alternative null statement must be true if the null hypoth- esis is false. The statistical notation for the alternative hypothesis is H1. It is the direct opposite of the null and is what the researcher hopes to be true, as it represents the research hypothesis. Unlike its opposite, the null hypothesis, it never contains the equal sign. H1 ≠ 0 is the statistical language used to represent the altervative hypothe- sis (Friedman, 2018). One question which is often asked is why we test for the null rather than the alternate hypothesis, which is the one we hope to prove. One of the reasons is that, in business and the social sciences, there are few things that are able to be determined 100 percent through statistical hypothesis testing. There are simply too many factors to consider establishing true causality. This is true with even tightly controlled experiments. Therefore, we assume there are other factors causing the out- come and we ﬁnd enough evidence against this to conclude that the alternate hypoth- esis must be a probable inﬂuence. Take the negative health effects on the use of tobacco products, for instance. It’s very difﬁcult to actually prove a link between the use of these products and ill health effects. However, evidence from tests and studies that have been conducted over the previous ﬁve decades would indicate otherwise. Enough evidence has been gathered to actually change the behavior of individuals, despite the fact that causality can never be established. This is the power of hypothesis testing (Friedman, 2018).

**Step 2: Choose the appropriate distribution**

For this step, the appropriate probability distribution is determined by whether you know the population standard deviation or need to rely on the sample standard devia- tion. Where the population standard deviation is known, one would use the z-value, which represents the standard normal distribution. For hypothesis tests where only the sample standard deviation is known, one would choose the t-distribution. It should be noted that, when sample sizes are large (i.e., 30 or more observations), both distribu- tions look the same and students are able to choose either (Friedman, 2018).

**Step 3: Ascertain the rejection and non-rejection regions**

In this step the signiﬁcance level, denoted as alpha, or α, is to be calculated. The signif- icance level represents the probability of rejecting the null hypothesis when it is true. For example, a signiﬁcance level of 0.01 indicates a one percent risk of concluding that a difference exists when there is no actual difference (Friedman, 2018).

**Step 4: Compute the value of the test statistics**

The calculation for the z-test statistic when the population standard deviation is known is

(5.1)

Where:

= Sample mean

sigma = Population standard deviation

n = Sample size

The rejection and non-rejection regions are separated by the values of the test statis- tic. The rejection region is the set of values for the test statistic that leads to rejection of H0. The non-rejection region is the set of values not in the rejection region, which lead to a non-rejection of H0 (Friedman, 2018).

The p-value is another quantitative measure for reporting the result of a hypothesis test. When the p-value is low, there is a greater likelihood of obtaining the same result. Therefore, a low p-value provides statistical evidence that the results of the test are not due to random sampling error alone. A p-value is equal to the chance of obtaining a test statistic equal to or more extreme, than the observed value of H0 (Friedman, 2018).

We then compare the p-value with α, there are three possible outcomes:

1. If the p-value < α, reject H0.
2. If the p-value >= α, do not reject H0.
3. If the p-value is low, then H0 must go.

The idea of hypothesis testing is to reject the null hypothesis if the sample data do not agree with the null hypothesis. Thus, if the observed test statistic is more weighted in the direction of the alternative hypothesis than is preferred from a decision risk per- spective, then one will be justiﬁed in rejecting the null hypothesis (Friedman, 2018).

**Step 5: Make a decision**

According to the results of your hypothesis test, you can ascertain whether your study fails to reject, accepts, or rejects the null hypothesis. It is important to note that the researcher should not accept the alternate hypothesis. This, as mentioned, is due to the fact that the hypothesis testing process is meant to gather evidence for a research hypothesis, as opposed to making a deﬁnitive conclusion. We also often discover that, when the results of a hypothesis test are reported in academic journal, it is common to ﬁnd that the researcher provides only the test statistic and its p-value, leaving it up to the reader to draw their own conclusions. This is also not a preferred practice; however, since the accuracy of precise p-values may be inﬂated (Friedman, 2018).

### One-Tailed versus Two-Tailed Hypothesis Tests

There are two considerations researchers need to make when conducting a hypothesis test: whether they are conducting a one-tailed or a two-tailed test. A two-tailed test involves rejecting the null hypothesis if the t-score for the sample is of a low probabil- ity in either direction. This test is justiﬁable when it is believed that the sample mean might differ from the hypothetical population mean, but we do not have good reason to expect the difference to go in any particular direction (McEvoy, 2018). An example of this would be wanting to compare the mean strength of auto parts from an original equipment manufacturer (OEM) parts supplier to a predetermined target value. In this case, we would want to use a two-tailed test because we are trying to ascertain whether the mean exceeds or is lower than the target value (McEvoy, 2018).

Conversely, we have the one-tailed test, with which we would only reject the null hypothesis if the t-score for the sample were a low probability in one direction that is speciﬁed before collecting the data. This test makes sense when we have good reason to expect that the sample mean will differ from the hypothetical population mean in a particular direction, which could be greater than or less than the hypothetical popula- tion mean. An example of this is that employee engagement scores would increase as a result of providing new incentives (McEvoy, 2018). When you have a one-tailed test you need to determine whether it is a right-tailed test or a left-tailed test, which are some- times also referred to as upper and lower tests. When one’s hypothesis has an inequal- ity that points to the right and contains a greater than (>) symbol, then this is a right- tailed test. For example, you might be comparing the life of refrigerator compressors before and after a manufacturing change (Glen, n.d.). If you want to know if the com- pressor life is greater than the original (e.g., 10 years), your hypothesis statements might be as follows:

* null hypothesis: No change or decrease observed (H0 ≤ 10).
* alternate hypothesis: Compressor life has increased (H1 > 10).

When one’s hypothesis has an inequality that points to the left and contains a less than (<) symbol, then this is a left-tailed test. For example, you might be comparing the length of time it takes to execute digital marketing campaigns before and after an implementation of customer relationship management software (Glen, n.d.). If you want to know if the cycle time for campaign execution is less than the original (let’s say three days), your hypothesis statements might be as follows:

* null hypothesis: No change or greater than (H0 ≥ 3).
* alternate hypothesis: Campaign execution time has decreased (H1 < 3).

It is important to note that whether a test is left- or right-tailed is determined by the alternate hypothesis (H1) and not the null hypothesis (Glen, n.d.).

Hypothesis tests are one of the greatest tools developed in the twentieth century. These tests are indirectly responsible for the development of vaccines, life-saving drugs (like the cure for Hepatitis-C), and doubling the life span of individuals in less than a century. Hypothesis tests are designed to determine whether two treatments or a single treatment has a signiﬁcant effect on the population, or whether any differen- ces are due to sampling error. All hypothesis tests start with a research hypothesis for the desired test element, followed by the creation of the null hypothesis, which assumes there is no statistically signiﬁcant difference between the population mean and the sample mean. If, after calculating the difference, the probability of the differ- ence happening by chance proves to be very low, then we have enough evidence to reject the null hypothesis, therefore supporting our alternate hypothesis related to our research hypothesis. The choice of which test to use depends on whether we know the sample standard deviation (t-test) or the population standard deviation (z-test).

## 5.2 Types of Hypothesis Testing When the Population Standard Deviation is Unknown

The t-test is used when the population standard deviation is unknown. The most important type of hypothesis testing in business and the social sciences, if we are focused on the differences between two means, is the t-test. In this section, we exam- ine three types of t-tests that are associated with speciﬁc research designs: the one- sample, dependent samples, and independent samples t-tests.

### One-Sample T-Test

The **one-sample t-test** is used when researchers want to compare a mean derived from a sample (M) with a hypothetical population mean (μ0). The null hypothesis is that the mean for the population (µ) is equal to the hypothetical population mean: μ = μ0. The alternative hypothesis is that the mean for the population is different from the hypothetical population mean: μ ≠ μ0 (McEvoy, 2018). To conduct this hypothesis test, we need to ascertain the probability of obtaining the sample mean if the null hypothesis were true. To do this we must first find the *p* value by computing a test statistic called *t* (McEvoy, 2018). The formula we use for *t* is as follows:

(5.2)

In this equation, top enclose x is the sample mean and (µ0 is the hypothetical population mean of interest), S2 is the sample standard deviation, and n is the sample size. If we calculate our t- score using a computer, we will receive both the t-score and the *p*-value as outputs. At this point, we would implement the ﬁnal step in the hypothesis testing process, i.e., make a decision based on the data. Assuming we decide to test the hypothesis at a 95 percent conﬁdence level, if *p* is less than .05, we reject the null hypothesis and con- clude that there is evidence that the population mean differs from the hypothetical mean. If *p* is greater than .05, we fail to reject the null hypothesis and conclude that there is not enough evidence to say that the population mean differs from the hypo- thetical mean (McEvoy, 2018).

If we want to calculate the t-score manually, we could use a table of **critical values** of t when alpha equals .05 to make our decision. Interestingly, these tables do not provide actual *p* values; instead, they provide the critical values of t for different degrees of freedom (df) when a is .05. There are also tables for different alpha levels, which vary depending on whether the test is one- or two-tailed. When implementing the decision making step, we consider any t-score beyond the critical value in either direction (for two-tailed tests only) to be in the most extreme ﬁve percent of t-scores when the null hypothesis is true and has a < value less than .05. Thus, if the t-score we compute is beyond the critical value in either direction, then we have enough evidence to reject the null hypothesis. However, if the t-score we compute is between the upper and lower critical values, then we fail to reject the null hypothesis (McEvoy, 2018).

**Critical values**

These are derived from a graph of a distribution that divides the graph into a reject region. If the t exceeds that value and enters the region, then the null hypothesis should be rejected.

### The Dependent Sample T-Test

The dependent sample t-test is often referred to as the paired t-test. This test is used to compare two means for the same sample tested at two different times or under two different conditions. An example of this in business is the comparison between a pre- test and posttest measuring knowledge of corporate ethics after employees take a new corporate ethics training module. The null hypothesis is that the means calculated from sample data at the two points in time (or under the two conditions) are the same in the population. The alternative hypothesis is that they are different. If there is good reason to expect that the difference will go in a particular direction (like in the previous example), then this test can be one-tailed. This test is appropriate for related samples since the same variation in both observations can be taken into account to make the test more powerful (McEvoy, 2018).

One can frame their interpretation of the dependent sample t-test as a particular type of one-sample t-test after the preparation process is completed. This process speciﬁ- cally involves subtracting the two scores for each participant to create a single, differ- ence score. With this ﬁrst step completed, the dependent sample t-test becomes a one- sample t-test using these difference scores. The hypothetical population mean ((0) is 0 because this is what the mean difference score would be if there were no difference (on average) between the two observations or two conditions of the same population. The null hypothesis can then be stated as the mean difference score in the population is 0 ((0 = 0), and the alternative hypothesis is stated as the mean difference score in the population is not 0 ((0 ≠ 0) (McEvoy, 2018).

### The Independent Sample T-Test

The third hypothesis testing method is the independent sample t-test, which is used to compare the means of two separate samples (Ma and Mb). The two samples might have been acquired via testing under different conditions in a between-subjects experiment, or they could be preexisting groups in a correlational study (e.g., married and single, or male and female). An example application of an independent sample t- test in business is if a random sample of customers were either handled through a live agent or a chatbot and we were to test whether there was a statistically signiﬁcant dif- ference in their mean satisfaction scores (McEvoy, 2018). The important thing to note about these samples is that they are chosen independently from one another. The null hypothesis is that the means of the two populations are the same ((1 = (2). The alter- native hypothesis is that they are not the same ((1 ≠ (2). This test should be designa- ted as one-tailed if the researcher has good reason to expect the difference in means to be either higher or lower (McEvoy, 2018). The 1 statistic in this test is considerably more complex due to the fact that were are now taking into account two sample sizes, two sample means, and two standard deviations (McEvoy, 2018). The formula is as fol- lows:

(5.3)

where

* mA and mB represent the mean value of the group A and B, respectively.
* nA and nB represent the sizes of the group A and B, respectively.
* s2 is an estimator of the pooled variance of the two groups.

It can be calculated as follows:

(5.4)

with degrees of freedom df=nA+nB−2.

Please note that the formula above contains the variances, which are the squared standard deviations contained inside the square root symbol. Also, lowercase n/ and n3 refer to the sample sizes in the two groups, which is the opposite of the symbol *N*, which, in statistical notation, generally refers to the total sample size. One ﬁnal thing to note is that there are N − 2 degrees of freedom for the independent sample t-test.

### Confidence Intervals for Hypothesis Testing

Another method of conducting hypothesis testing is through interval estimation, specif- ically conﬁdence intervals. Curran-Everett (2009) discusses the fact that, unlike hypoth- esis tests, conﬁdence intervals are a fairly recent phenomenon, developed by Jerzy Neyman in the 1930s. Both the hypothesis test and conﬁdence intervals can accomplish the same thing, but are often used for different purposes. For example, many research- ers will use hypothesis testing when they have a pre-speciﬁed hypothesis and signiﬁ- cance level and want to do a strict comparison. Conversely, when they want to describe the magnitude of an effect (e.g., mean difference or odds ratio), or a single sample, a conﬁdence interval may be more useful (Curran-Everett, 2009).

Which test to use depends on whether we know the sample standard deviation (t-test) or the population standard deviation (z-test). There are also different types of t-tests, which are chosen based on whether there is a single sample or two samples; for the latter, we must also determine whether the samples are related or independent. One ﬁnal note is that hypothesis testing is not just for means, but can also be used to test other statistics, including proportions.

### Summary

Hypothesis tests are indirectly responsible for the development of vaccines, life- saving drugs, and doubling the life span of individuals in less than a century.Hypothesis tests are one of the greatest tools developed in the twentieth century. The premise of the hypothesis test is simple; they determine whether two treat- ments or a single treatment has a signiﬁcant effect over the population, or whether any differences are due to sampling error. All hypothesis tests start with a research hypothesis, followed by the creation of the null hypothesis, which assumes there is no statistically signiﬁcant difference between the population mean and the sample mean. If, after calculating the difference, the probability of the difference happen- ing by chance proves to be very low, then we then have enough evidence to reject the null hypothesis, which then supports our alternate hypothesis related to our research hypothesis.

Which test to use depends on whether we know the sample standard deviation (t- test) or the population standard deviation (z-test). This distinction becomes less important with larger sample sizes. There are also different types of t-tests, which are chosen based on whether there is a single sample or two samples. Many researchers will use traditional hypothesis testing when they have a pre-speciﬁed hypothesis and signiﬁcance level and want to do a strict comparison. Hypothesis testing is not just for means, but can also be used to test other statistics, including proportions when the data are normal, and statistics, such as medians with non- parametric tests in which the data are not normally distributed.

## 6. Simple Linear Regression

## **Study Goals**

On completion of this unit, you will be able to …

… identify and understand the applications for practicing simple linear regression analysis.

… design and compute a simple linear regression analysis.

… interpret and validate a simple linear regression model.

## Case Study

The supermarket chain, Anderson’s, wants to ascertain whether advertising expendi- tures for targeted campaigns over the previous three years have had an impact on sales. The chief marketing ofﬁcer Alice Taylor elicits the assistance of Helen Weber, the chief digital ofﬁcer, to determine whether Anderson’s is able to attribute their ad spend to increased sales. Helen explains to Alice that it is indeed possible, and actually quite common for ﬁrms who spend a good portion of their budgets on advertising, to be able to measure the results of the campaigns. She goes on to explain that a popular method of measure in the efﬁcacy of the advertising effort of ﬁrms is to build attribution mod- els using simple linear regression models.

As a next step, Helen agrees to collect data related to the previous 36 months of adver- tising spend and company sales. The advertising spend serves as the independent vari- able and the gross sales serves as the dependent variable in this study. Helen plans to run the regression analysis in Excel and then share the results with Alice and the branding and advertising team. Alice anxiously awaits the results and hopes that there will be a strong positive relationship between their advertising efforts and sales so she can present the results to the executive team to gain additional funds to help grow the business.

## 6.1 Simple Linear Regression—Concepts, Approach, and Quality Assessment

Karl Pearson developed a mathematical rigor around the Pearson product moment cor- relation, and was inspired by Sir Francis Galton, the accomplished nineteenth century British scientist, who was originally thought to have conceived both correlation and regression analysis (Stanton, 2001). In an era of science focused on proving or disprov- ing the theory of evolution, Galton’s study of genetics sparked the idea that led to the development of linear regression.

In a famous experiment with sweet peas, Galton gained the idea for regression by plot- ting a scatter diagram of the sizes of daughter peas in relation to the sizes of their mother (sweet peas are self-fertilizing). While providing evidence for heredity, it also gave Galton the idea that two variables which were related (i.e., the size of mother and daughter sweet peas) could also be predicted. The generalized form of regression anal- ysis was developed later by numerous data scientists (Stanton, 2001).

Other contributors to regression analysis include Carl Friedrich Gauss and Adrien-Marie Legendre, scientists who both independently discovered an essential feature of regres- sion analysis: the method of least squares. Least squares is a statistical optimization technique in which the sum of the squared errors are minimized. Both of these scien- tists used this method to understand the orbits of celestial bodies (Stigler, 1981). While the applications of linear regression increased dramatically over the next 150 years, very little advancement to the technique was made, primarily due to the lack of high- speed computing (Hocking, 1983). More recent developments in the techniques of regression analysis, aided by increased computing power, include contributions from statisticians John Tukey, George Box, David Cox, Sanford Weisberg, and Maurice Kendall, and, even more recently, David Freedman and Andrew Gelman (Hocking, 1983).

## Generalized Approach to Regression

Almost all statistical programming languages from, R to SAS and even spreadsheet pro- grams such as Excel, make the calculation of regression analysis seem relatively easy while having linear algebra operating as their hidden engine powering the regression equations. We will brieﬂy discuss linear algebra, as it is necessary for an understanding of the linear algebraic origins of **linear regression**.

In its simplest form, appropriately referred to as simple linear regression, our goal is to ﬁnd the best ﬁt line (from which we can then predict with new observations) through a set of data points: (x1, y1), (x2, y2), … (xn, yn) (Sundaram, 2020). For example, this could be used to try to predict employee satisfaction scores as a function of average employee tenure. But the real question to ask is what does the best ﬁt mean? Keeping in mind that the equation for a straight line is y = mx + b, where y is our dependent variable, x is our independent variable, m is the slope of the line, and b (c in the UK) is the constant and also referred to as the y-intercept. Remember, the constant is y when x = 0. If we can ﬁnd a slope and an intercept for a single line that passes through all the possible data points, then it would certainly be considered the best ﬁt line, simply because it is a perfect ﬁt. In terms of regression, there would be no errors (Sundaram, 2020).

**Linear regression**

This describes a ser- ies of modeling techniques, which can have different error distributions, but all relate in some way to simple linear regression. They are known as general linear mod- els (GLM).

However, in the majority of cases, such a perfect line does not exist. Therefore, we default to searching for another line that measures the error of each data point when a connecting line is drawn parallel to the y-axis, from the data points to the regression line. This should ensure that the sum of all such errors should be minimized. Though it sounds quite simple, we deploy matrix and linear algebra behind the scenes to make this happen (Sundaram, 2020).

Regression, paired with hypothesis testing, are the key value drivers of statistical analy- sis. However, despite all of its popular applications, from credit scoring to sales fore- casting, regression should not be considered a single entity. Rather, regression is a vari- ety of statistical methods, all with one fundamental idea in common:

*Dependent Variable (y) = Constant (b) + Slope (m)\*Independent Variable (x) + Error (e)*

The dependent variable in the simple linear regression equation, also known as the outcome variable, is something the researcher will want to predict or explain. For example, in a customer management context we may want to ﬁnd a customer satisfac- tion score measured on a zero to ten rating scale. The independent variable in a simple linear regression equation, which is also known as the predictor variable, is what is used to explain or predict the dependent variable. Continuing with examples of cus- tomer management, we could assess the length of a customer service call in minutes (Gray & Gray, 2017). The constant term in the equation above as previously stated is also referred to as the Y intercept. The y represents the dependent or the outcome var- iable, and x is the independent variable. The slope shows amount Y changes when X changes by a speciﬁc amount (Gray & Gray, 2017).

The error term is a critical one and constantly reminds us that it's virtually impossible to predict y from x with 100 percent certainty. Sometimes no errors occur in a model, however, this should raise concerns that there may be something wrong with the model. Analysts should try to continually reduce the errors and improve the model (Gray & Gray, 2017).

An introduction of simple linear regression is necessary to be able to conceptualize the process. However, in many business applications, we have more than one independent variable and, in this case, we can utilize a more advanced from of regression analysis known as **multiple regression** (Gray & Gray, 2017).

**Multiple regression**

Solving multiple regression problems by hand requires one to think of the points on an n- dimensional surface in space, with n rep- resenting the num- ber of independent variables. The shapeof this surface depends on the model structure.

One of the key assumptions of simple linear regression is a linear relationship between the two variables. However, some business applications deal with nonlinear relation- ships. In these cases, similar to multiple regression, the regression class of algorithms also covers nonlinear relationships with more advanced models like polynomial regres- sion and Gompertz regression. Even taking these special algorithms into account, many linear regression models are, in fact, linear only after transformations of the variables have been made (Gray & Gray, 2017).

Some of the other considerations related to regression analysis are related to the type of dependent variable we are working with. Simple linear regression assumes a contin- uous dependent variable at the interval or ratio level of measurement. Sometimes, researchers want to predict an outcome which is dichotomous or binary. Loan default on a credit score is an example of such a variable, since it is measuring whether a cus- tomer repays their loan or does not. Because a loan default variable is a binary varia- ble and not a continuous numeric, using a simple linear regression that assumes a numeric dependent variable would not be an acceptable statistic. Logistic or probit regression in this case would be the correct choice (Gray & Gray, 2017). Moreover, if the dependent variable is categorical but has three or more categories, multinomial logistic regression (based on a multinomial distribution) is used. An application of regression where there are often three or more categories to be modeled is in conjoint analysis, in which respondents typically choose from three or more products in each choice task (Gray & Gray, 2017).

Sometimes, we have a categorical dependent variable in which the categories are ordered, e.g., light, medium, and dark colors. In this case one would choose ordinal logistic or probit regression (Gray & Gray, 2017). Finally, when researchers have a dependent variable, which is count data (i.e., how often a customer visited the store in one year), Poisson and negative binomial regression are two of the best choices for modeling said data (Gray & Gray, 2017). These examples are just some of the most pop- ular regression techniques. Others include quantile regression, box-cox regression, truncated and censored regression, hurdle regression, and nonparametric regression, as well as regression methods for time series data (Gray & Gray, 2017).

Developing a Linear Regression Model

As mentioned, linear regression is a statistical tool used to measure and predict the relationship between two variables: one independent (the explanatory variable) and one dependent (the predictor variable). This is done by ﬁtting a linear equation to observed data. For example, a business analyst might want to model and relate the annual purchases of customers to their income levels using a simple linear regression model (Yale University, 1997). Before even attempting to ﬁt a simple linear regression model to the data that have been collected, an analyst should ﬁrst look for some type of evidence to determine whether or not there is a relationship between the variables of interest. They should always start with a strong hunch or research hypothesis that it would be logical for a relationship to exist between two variables (Yale University, 1997).

When trying to ascertain the relationship between variables in advance of conducting a regression analysis, it’s critically important to recognize that a correlation between two variables does not necessarily assume or imply that one variable causes the other. For example, higher scores on a pre-employment test do not necessarily cause better employee performance. However, this does not take away from the fact that there is some signiﬁcant association between the two variables.

Dr. William Edwards Deming, the famous statistician and engineer, and one of the fathers of modern quality, often shared the idea that graphs are friends that help us to understand data. On that same note, a scatterplot of scatter diagram should be one of the ﬁrst tools used to ascertain the strength of the relationship between two variables. If there appears to be a lack of association (randomness) between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing linear trends by plotting the x and y variables), then ﬁtting a linear regression model to the data will probably not be a productive use of time.

In addition to creating and analyzing a scatter diagram, a researcher at this stage of the modeling process can also calculate a Pearson product moment correlation coefﬁcient, which is a numerical modeling of the association between two variables. The output of this model yields what is known as the correlation coefﬁcient, which is a value between-1 and 1, indicating the strength and direction of the association of the observed data for the two variables. A positive coefﬁcient indicates that, as the x variable increases, so does the y variable. A negative correlation indicates that as the x variable in increa- ses, the y variable, in turn, decreases (Yale University, 1997). The form of the simple lin- ear regression equation is B = 4 + CD, where D represents the explanatory variable and B represents the dependent variable. The slope of the line is represented by C, and the intercept is represented by a (remember a represents the value of y when x = 0) (Yale University, 1997).

Quality Assessment for Simple Linear Regression

There are several ways to assess the quality of a regression equation. The ﬁrst is a numeric method, and the second is a graphical method, which involves plotting the actual versus predicted values from the regression analysis. The numeric method involves the Pearson’s correlation coefﬁcient, which is a statistical tool that ranges from 0 to +1 and -1 and describes a linear relationship between two variables. In the case of simple linear regression, we use the coefﬁcient of determination r2, which is a measure of how well the regression model describes the observed data. In simple lin- ear regression analysis, r2 is simply the Pearson’s correlation coefﬁcient squared. The interpretation of this statistic is that it represents the percentage of variance in the dependent variable by knowing the independent variable. Of course, the higher the percentage, the better, as long as the model does not overﬁt. Overﬁtting is when a model is only performing on the sample from which it was developed, and cannot extrapolate consistently and effectively when new samples are introduced (Schneider et al., 2010).

One can also test the regression coefﬁcients for statistical signiﬁcance. The null hypothesis for a simple linear regression is b = 0, meaning there is no relationship between variables, and the regression coefﬁcient is therefore zero. This can then be tested with a t-test. A 95 percent conﬁdence interval for the regression coefﬁcient can also be computed (Schneider et al., 2010).

Regression analysis in both the simple linear regression form and its more sophistica- ted nonlinear and multiple regression counterparts are considered one of the most important statistical procedures in business and our personal lives today. Not only can this tool help researchers to ascertain whether or not on or more variables have a stat- istical relationship, but it can also be used to predict the future when new observations are made in the independent variable. This module introduced the concept of simple linear regression so that students can conceptualize and easily learn the process. The equation for simple linear regression follows the form B = 4 + CD, where D represents the explanatory variable and B represents the dependent variable. The slope of the line is represented by C, and the intercept is represented by 4 (Yale University, 1997). The idea is to used the least-squares algorithm to create a line across all points between the two variables which minimizes the sum of the squared errors. In the case if simple linear regression we use the coefﬁcient of determination r2, which is a meas- ure of how well the regression model describes the observed data to assess the quality of our regression analysis. In simple linear regression analysis, r2 is simply the Pear- son’s correlation coefﬁcient squared. The interpretation of this statistic is that it repre- sents the percentage of variance in the dependent variable there is by knowing the independent variable

6.2 Applications of Simple Least-Squares Regression

The least-squares is the most common method for fitting a regression. This method minimizes the sum of the squared errors. What this means is that a best-fitting line is created for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (the vertical deviation is zero if a point lies on the fitted line exactly). The squaring operation means the deviations are first squared, then summed, the positive and negative values do not cancel each other out (Yale University, 1997).

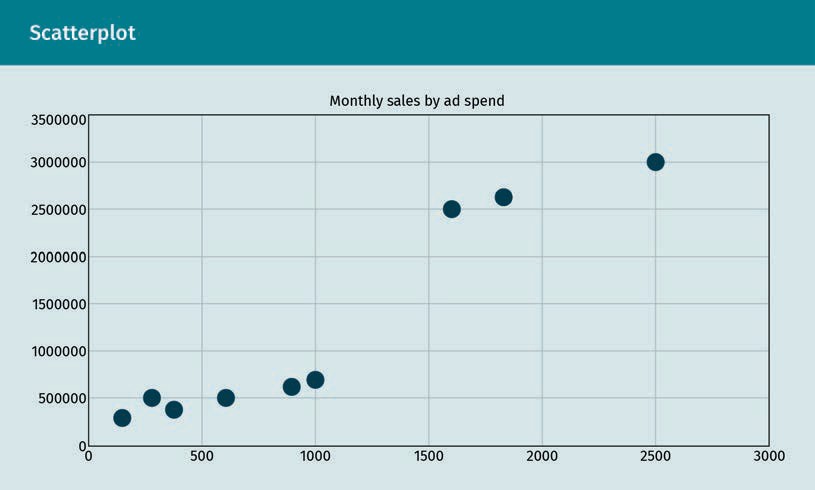
Example

This example will teach students how to run a simple linear regression analy- sis in Excel and how to interpret the summary output. Below are the sample data. The research question is as follows: Is there a relation between monthly sales (dependent variable) and monthly advertising spend (independent variable). Moreover, can we pre- dict sales if we know the advertising spend?

|  |  |
| --- | --- |
| Sample Advertising Data | |
| Monthly sales | Monthly ad spend |
| 3,000,000 | 2,500 |
| 500,000 | 300 |
| 300,000 | 200 |
| 400,000 | 400 |
| 500,000 | 600 |
| 2,000,000 | 1,600 |
| 2,100,000 | 1,800 |
| 600,000 | 900 |
| 700,000 | 1,000 |

### Scatterplot

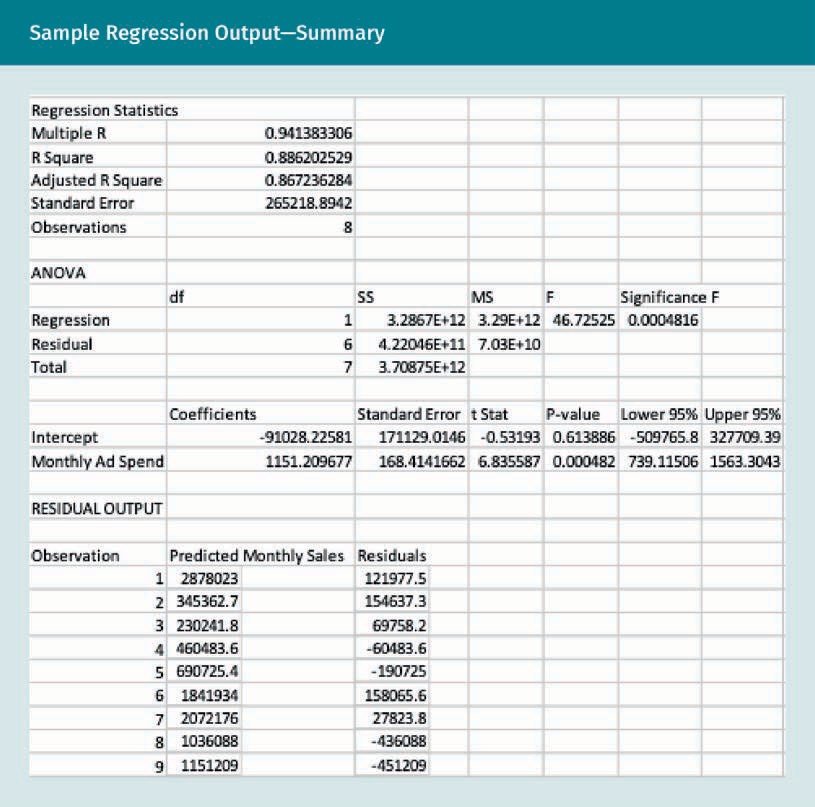
Before we begin the process of building a simple linear regression model, it makes sense to ﬁrst create a scatterplot to assess the extent of the linear relationship between the two variables (Excel Easy, n.d.). An example of such a scatterplot is shown below.



From the scatterplot above, we can see a fairly strong positive correlation between the two variables, which infers that, as advertising spend is increased, we will also observe an increase in monthly sales (Excel Easy, n.d.).

### Simple Linear Regression

Next, we run a regression analysis on this dataset in Excel using the Analysis ToolPak add-in to the Data Analysis Tab. An example of how the total output would be repre- sented in Excel is given in the ﬁgure below.



R-Square

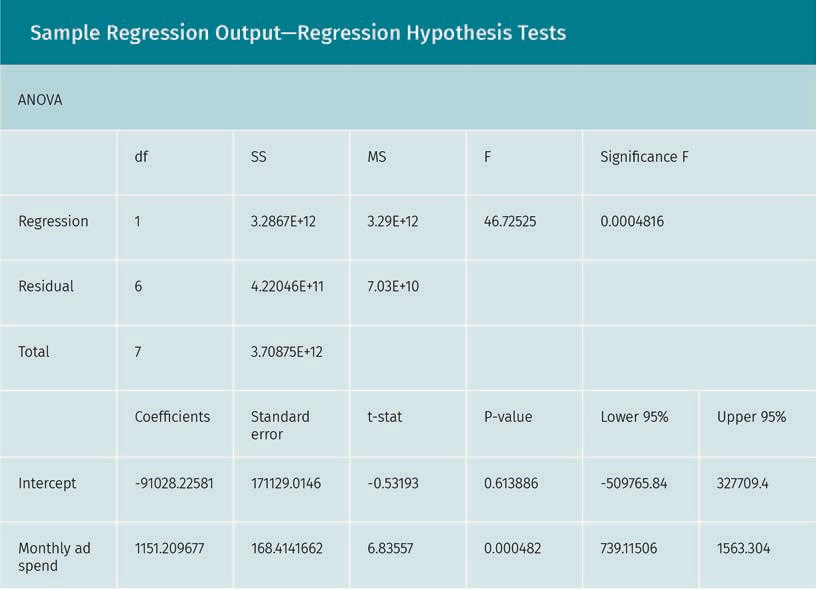
The ﬁrst thing we check on this analysis is the r-square in the summary output. R- square is the correlation coefﬁcient squared. The r-square, or 2², value in this regres- sion analysis is 0.886 (the square of the correlation coefﬁcient 0.941), indicating that 88.6 percent of the variation in one variable (i.e., monthly sales) may be explained by the other (i.e., advertising spend) (Excel Easy, n.d.).

|  |  |
| --- | --- |
| Sample Regression Output: Regression Statistics | |
| Summary output | |
| Multiple R | 0.941383306 |
| R-square | 0.886202529 |
| Adjusted R-square | 0.867236284 |
| Standard error | 265218.8942 |
| Observations | 8 |

### Significance F and P-values

Next, w need to ascertain whether the results are reliable and statistically signiﬁcant. Take a look at signiﬁcance F (0.000481). If this value is less than 0.05, then you can be generally conﬁdent in your analysis. If you discover that signiﬁcance F is greater than 0.05, you probably need to search for a new independent variable and rerun your regression analysis. This should be repeated until signiﬁcance F drops below 0.05.

Most, or all, p-values should be below 0.05. In our example, this is the case with monthly ad spend (0.000482). Our intercept is not statistically signiﬁcant because there isn’t sufﬁcient statistical evidence that it is different from zero (0.61388). However, it seems reasonable that, without advertising, there would be no sales, so we will keep this regression analysis (Excel Easy, n.d.).



### Coefficients

The regression line is y = monthly sales = -91028.225 + 1151.209 · monthly ad spend. In other words, for each unit increase in ad spend, sales increases by 1,151.209. This is val- uable information. You can now also create forecasts with this simple linear regression model. For example, if monthly ad spend equals 2657, you might be able to achieve a quantity sold of 0 + 1151.209 · 2657 = 3,058,762.31 (Excel Easy, n.d.).

### Residuals

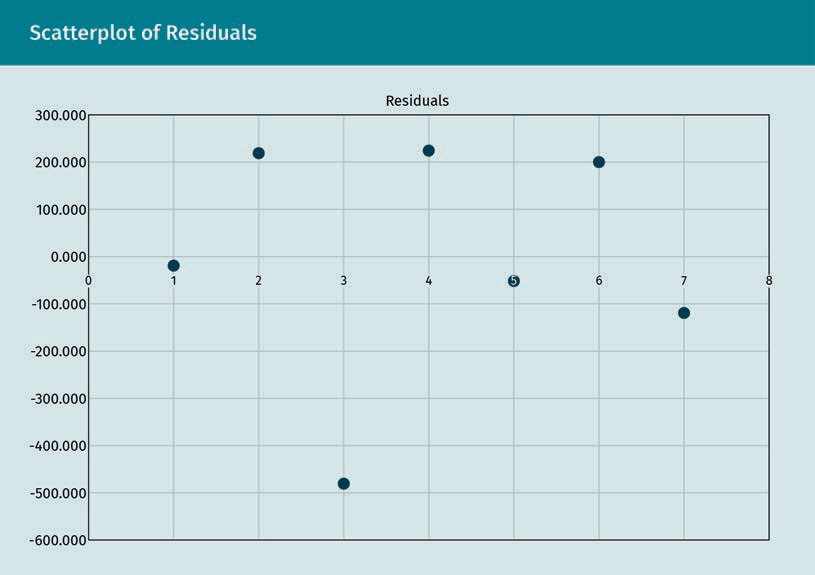
An analysis of the residuals can help us ascertain how far away the predicted data points are from the actual data points by using the given regression equation. For illus- trative purposes, the ﬁrst data point in our advertising spend example equals 2500. Using the equation, the predicted data point equals 0 + 1151.209 · 2500 = 2,878,023, giving a residual of 3,000,000 - 2,878,023 = 121,977 (Excel Easy, n.d.).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Observation | Predicted monthly sales | | Residuals | |
| 1 | 2878023 |  | 121977.5 |  |
| 2 | 345362.7 |  | 154637.3 |  |
| 3 | 230241.8 |  | 69758.2 |  |
| 4 | 460483.6 |  | -60483.6 |  |
| 5 | 690725.4 |  | -190725 |  |
| 6 | 1841934 |  | 158065.6 |  |
| 7 | 2072176 |  | 27823.8 |  |
| 8 | 1036088 |  | -436088 |  |
| 9 | 1151209 |  | -451209 |  |

You can also create a scatter plot of these **residuals**.

**Residual**

When a residual plot shows a random pat- tern, it suggests a good ﬁt for a linear model. If the pat- terns in the residual plot are non-random (i.e., U-shaped and inverted U-shaped), this suggests a bet- ter ﬁt for a nonlinear model.



Outliers and Influential Observations

A point that lies far from a regression line that has been computed for a group of data, such as in the group shown above, is often referred to as an outlier or extreme value. It is important that any outlier be addressed, since these may result in a ill-ﬁtting regres- sion. The plot of the regression line for the advertising sample (created by highlighting the data and creating a scatterplot with a linear trendline option in Excel) is shown below.



Please note that points that lie far from the other data in the horizontal direction are referred to as inﬂuential observations. We make the distinction between inﬂuential observations and outliers because the former may have a signiﬁcant impact on the slope of the regression line. In the case of the data in this application, most of the observations tend to be close to the prediction, so we do not see any evidence of outli- ers or inﬂuential observations being present (Yale University, 1997).

Lurking or Confounding Variables

Whenever the relationship between two variables is signiﬁcantly affected by the pres- ence of a third variable that has not been included in the modeling effort, a lurking or confounding variable is present. The presence of lurking variables may manifest in the creation of nonlinear trends, which are visible in the relationship between an explana- tory and dependent variable (Yale University, 1997).

Extrapolation

The actual range of the data should always be carefully considered whenever a linear regression model is ﬁt to a dataset. It is inappropriate to attempt to use a regression equation to predict values outside of this range, as it may yield unreliable answers. This practice is known as extrapolation. One example of this is a linear model that relates employee health risk scores to age for younger employees in their twenties and thir- ties. Applying such a model to middle-aged employees in their forties and sixties would not be appropriate, since the relationship between age and employee health risk scores is not consistent for all age groups (Yale University, 1997).

### Simple Linear Regression Assumptions

When conducting a linear regression, certain assumptions must be upheld. Each of these assumptions may have higher or lower importance depending on the data and application. There are three key assumptions (Yale University, 1997):

1. Homogeneity of variance (homoscedasticity). With homoscedasticity, we assume that the size of the error in our prediction doesn’t vary signiﬁcantly across the values of the independent variable.
2. Independence of observations. With independent observations, we assume that there are no hidden relationships among the observations and that all of the obser- vations in the dataset were gathered using statistically valid sampling methods.
3. Normality. Normality assumes that the data are following a pattern that closely approximates the normal distribution.

Linear regression makes the additional assumption that the relationship between the dependent and independent variable is linear. The line of best ﬁt through the data points is a straight line, rather than a curve, which makes the relationship nonlinear or curvilinear. Please note that if your data do not meet the assumptions one through three above, you may be able to use a nonparametric regression, such as kernel analy- sis or regression trees. For violation of the assumption of linear regression, the nonlin- ear regression methods, including polynomial and Gompertz regression, may be poten- tial solutions (Yale University, 1997).

The unique property created by regression analysis lends it to useful applications including, but not limited to, marketing and operations forecasting, credit scoring, investment portfolio allocation, insurance underwriting, and pricing optimization. For simple linear regression, it is recommended to use the least squares algorithm, which is the most common method for ﬁtting a regression equation. This method minimizes the sum of the squared errors. This means that a best-ﬁtting line is created for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (the vertical deviation is zero if a point lies directly on the ﬁtted line). The least squares algorithm was used to create a line across all points between the two variables, which minimizes the sum of the squared errors for the mar- keting application we developed. Researchers must remember to always follow several key assumptions when conducting a regression analysis.

### Summary

Regression analysis is helpful to researchers seeking to ascertain whether one or more variables have a statistical relationship, as well as to predict the outcomes when new observations are made about the independent variable. For these rea- sons, this tool is considered one of the most important statistical procedures in business and our personal lives today. The equation for simple linear regression follows the form y = a + bX, where x represents the explanatory variable and y represents the dependent variable. The slope of the line is represented by b, and the intercept is represented by a (Yale University, 1997).

The least squares algorithm is an optimization procedure utilized to create a line across all points between the two variables, which minimizes the sum of the squared errors. We then use the coefﬁcient of determination r2 to assess the qual- ity of our regression analysis. We interpret r2 as the percentage of variance in the dependent variable, knowing the independent variable.

This unit demonstrated how simple linear regression works in a marketing applica- tion, but additional applications of regression analysis exist, including other mar- keting applications, sales and operations forecasting, pricing optimization, credit scoring, investment portfolio allocation, and insurance underwriting. An analysis of the residuals should always be considered when conducting a regression analysis in order to help ascertain, by using the regression equation above, how far away the predicted data points are from the actual data points. Researchers must be aware of the presence of inﬂuential observations and outliers, because these can both have an impact on the regression analysis. The presence of lurking variables may also create unwanted nonlinear trends when conducting a regression analysis. Finally, there are four key assumptions that researchers must remember to follow when conducting a regression analysis. If these assumptions are violated, it could lead to biased or unreliable results when interpreting the regression analysis.

# Appendix 1 – References

Adhikari, A., & Pitman, J. (2020). *Probability for data science*. Prob 140. <http://prob140.org/fa18/textbook/chapters/Chapter_24/01_Bivariate_Normal_Distribution>

Ahmad, F. (2016). *Basic statistical analysis for research: Step by step analysis*. Penerbit Universiti Tun Hussein Onn Malaysia (UTHM).

Akoglu, H. (2018). User’s guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91—93. <https://doi-org.pxz.iubh.de:8443/10.1016/j.tjem.2018.08.001>

Altman D., G, & Bland J., M. (2011). Statistics notes: How to obtain the confidence interval from a P value. *BMJ: British Medical Journal*, *343*(7825), 681. <https://doi.org/10.1136/bmj.d2090>

Amadeo, K. (2021). What are the odds of winning the lottery? How to improve your luck. *The Balance*. https://[www.thebalance.com/what-are-the-odds-of-winning-the-lot-](http://www.thebalance.com/what-are-the-odds-of-winning-the-lot-) tery-3306232

Attenborough, M. (2003), *Mathematics for electrical engineering and computing.* Elsevier.

Bandhari, P. (2020, November 5). The standard normal distribution. *Scribbr*. <https://www.scribbr.com/statistics/standard-normal-distribution/>

Britannica. (n.d.). *Dice.* https://[www.britannica.com/topic/dice](http://www.britannica.com/topic/dice)

Byjus Classes. (2021). *Probability and statistics symbols*. [<https://byjus.com/maths/probability-and-statistics-symbols/>](https://byjus.com/maths/probability-and-statistics-symbols/)

Center for Innovation in Teaching and Learning (CITL). (n.d.). *Contents of nomenclature for the fourfold (2x2) contingency table.* Central Illinois University. <https://cdn.citl.illinois.edu/courses/KINES401/Lesson4_lectures/Lecture4_P4/web_data/file2contents.htm>

Crooks, G. (2019). *Field guide to continuous probability distributions*. Berkeley Institute for Theoretical Science.

Curran-Everett, D. (2009). Explorations in statistics: Confidence intervals. *Advances in Physiology Education*, 33(2), 87—90.

Di Paola, G., Bertani, A., De Monte, L., & Tuzzolino, F. (2018, February). A brief introduction to probability. *Journal of Thoracic Disease*, *10*(2), 1129—1132. <https://jtd.amegroups.com/article/view/18920/15055>

Doane, D. P. (2016). *Applied statistics in business and economics* (5th ed.). McGraw-Hill.

Excel Easy (n.d.) *Regression*. <https://www.excel-easy.com/examples/regression.html>

Farber, B., & Larson, R. (2017, November 8). Z-0 normal distribution table. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:TABELA_DE_DISTRIBUI%C3%87%C3%83O_NORMAL_DE_Z-0.png>

Fogarty, D. J. (2017). *From beer to derivatives: The history of analytics in business*.Sara Book.

Friedman, A. (2018). *Statistics for LIS with open-source R*. Rowman & Littlefield.

Frost, J. (2019). *Introduction to Statistics: An intuitive guide for analyzing data and unlocking discoveries*. Statistics by Jim Publishing.

Glen, S. (n.d.). One tailed test or two in hypothesis testing; One tailed distribution area. *Statistics How To*. <https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/>

Gray, K., & Gray, C. (2017). *Regression analysis: A primer*. KDnuggets. <https://www.kdnuggets.com/2017/02/regression-analysis-primer.html>

Greenfield, M. L., Kuhn, J. E., & Wojtys, E. M. (1998). A statistics primer: Confidence intervals. *The American Journal of Sports Medicine*, *26*(1), 145—149.

Hayes, D. (2011, December 8). Lab2 histogram of MPG1000. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Lab2_Histogram_of_MPG1000.gif>

Hocking, R. (1983). Developments in linear regression methodology: 1959—1982. *Technometrics*, *25*(3), 219—230. <https://doi.org/10.2307/1268603>

Knapp, T. R. (1980). *Regression toward the mean. Statistics. [and] Basic descriptive statistics. Descriptive statistics. [and] Approximations in probability calculations. Applications of statistics. Modules and monographs in undergraduate mathematics and its applications project. UMAP Units 406, 426, 443*. Education Development Center. <https://files.eric.ed.gov/fulltext/ED218128.pdf>

Laerd Statistics. (2019, March 3). Pearson correlation coefficient and associated scatterplots. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Pearson_Correlation_Coefficient_and_associated_scatterplots.png>

[Lehmann, E. L.](https://en.wikipedia.org/wiki/Erich_Leo_Lehmann), & Casella, G. (1998). *Theory of point estimation* (2nd ed.). Springer.

Macfarlane, G. (2016, February 4). Population time series graph. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Population_time_series_graph.jpg>

Magiya, J. (2019). Kendall rank correlation explained. Towards Data Science. *Medium*. <https://towardsdatascience.com/kendall-rank-correlation-explained-dee01d99c535>

McEvoy, D. M. (2018). *A Guide to Business Statistics*. Wiley & Sons.

Microsoft. (n.d.). *Statistical functions (reference).* <https://support.microsoft.com/en-us/office/statistical-functions-reference-624dac86-a375-4435-bc25-76d659719ffd>

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Freeman and Co.

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: Part 14 of a series on evaluation of scientific publications. *Deutsches Arzteblatt International, 107*(4), 776—782. <https://www.aerzteblatt.de/int/archive/article/79009>

Shakil, M., Singh, J. N., & Golam Kibria, B. M. (2010). On a family of product distributions based on the Whittaker functions and generalized Pearson differential equation. *Pakistan Journal of Statistics* 26(1).

Skbkekas. (2010). Student t pdf. *Wikimedia Commons.* <https://commons.wikimedia.org/wiki/File:Student_t_pdf.svg>

Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9(3). <https://doi.org/10.1080/10691898.2001.11910537>

Stigler, S. (1981). Gauss and the invention of least squares. *Annals of Statistics*, 9(3), 465—474.

Sundaram, G. (2020). *Building Linear Regression (Least Squares) with Linear Algebra* Towards Data Science. <https://towardsdatascience.com/building-linear-regression-least-squares-with-linear-algebra-2adf071dd5dd>

Tattar, N. P., Ramaiah, S., &Manjunath, B. G. (2016). *A Course in Statistics with R*. Wiley & Sons.

Toews, M. W. (2007, April 7). Standard deviation diagram. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg>

Toews, M. W. (2007, July 9). Kernel density. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Kernel_density.svg>

Viti, A., Terzi, A., & Bertolaccini, L. (2015). A practical overview on probability distributions. *Journal of Thoracic Disease*, *7*(3), E7–E10. <https://doi.org/10.3978/j.issn.2072-1439.2015.01.37>

Viniciuslima94. (2017, June 29). Fig2 tipaper. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Fig2_tipaper.png>

Virginia Tech. (1999). *The t-distribution and its use in hypothesis testing*. https://simon.cs.vt.edu/SoSci/converted/T-Dist/

Yale University (1997) *Statistics department: Linear regression course*. <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>

Yemelyanov, D. (2020). Learn heteroskedasticity in 2 minutes. *Medium*. <https://medium.com/riga-data-science-club/learn-heteroskedasticity-in-1-minute-42f678911389>

# Appendix 2 – List of Tables and Figures

**Time Series Graph**

Source: Macfarlane (2016). [CC BY-SA 4.0]

--------------------------------------------------------------------------------------

**Histogram**

Source: Hayes (2011). [CC BY-SA 3.0]

--------------------------------------------------------------------------------------

**Standard Normal Distribution**

Source: Bhandari (2020).

--------------------------------------------------------------------------------------

**Standard Normal Distribution Probabilities**

Source: Toews (2007, April). [CC BY 2.5]

--------------------------------------------------------------------------------------

**Areas Under the Normal Curve Example**

Source: Farber & Larson (2017). [CC BY-SA 4.0]

--------------------------------------------------------------------------------------

**2 x 2 Crosstabulation**

Source: CITL (n.d.).

--------------------------------------------------------------------------------------

**Homoscedasticity and Heteroscedasticity**

Source: Yemelyanov (2020).

--------------------------------------------------------------------------------------

**Standard Bivariate Normal Distribution**

Source: Adhikari & Pitman (2020).

--------------------------------------------------------------------------------------

**Correlation Coefficient Examples**

Source: Laerd Statistics (2019). [CC BY-SA 4.0]

--------------------------------------------------------------------------------------

**Monotonic vs. Non-Monotonic Relationships**

Source: Magiya (2019).

--------------------------------------------------------------------------

**Discrete Probability Distribution**

Source: Viniciuslima94 (2017). [CC BY-SA 4.0]

--------------------------------------------------------------------------

**Continuous Probability Distribution**

Source: Toews (2007, July). [CC BY-SA 4.0]

--------------------------------------------------------------------------

**Typical Normal Distribution**

Source: Toews (2007, April). [CC BY 2.5]

--------------------------------------------------------------------------

**Student’s t-Distribution**

Source: Skbkekas (2010). [CC BY 3.0]

--------------------------------------------------------------------------

**All other tables and figures**

Source: David Fogarty (2021).