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

# Learning Objectives

The world we live in contains a large amount of data which is collected every hour every minute and every second. Data from our phones, our cars, our computers and even our own bodies are constantly being uploaded to servers. This data being collected has insights embedded within that challenge our abilities as humans to extract without the special tools available to us through the study of statistics. Statistics is a special branch of mathematics which addresses the collection, analysis 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 address the following topics

▪ understand the relevance of data to answer empirical questions.

▪ apply statistical methods in the overall context of concrete problems.

▪ solve statistical problems by using special statistical software.

Statistical analysis (statistics) is an important tool used by business managers to make decisions under uncertainty. A statistic is a single number used to explain a set of data. This course will prepare students to be able to learn 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.

… attain knowledge on how to extract insights from a sample using inferential statistics.

… understand how to compute and interpret probabilities.

# 1. Basics

## Case Study

The German-based company TechnoBank is a financial services company that has faced increased competition over the last few years and desired to gather data on its customers to better understand customer sentiments and loyalty. The marketing team, under the leadership of the chief marketing officer (CMO) Kristin Gehring, decides to design a questionnaire and conduct a sample of current and lapsed customers. Kristen asks the market research lead, Stefan Reidling to lead the project. Stefan outlines the process of questionnaire design and sampling from the Technobank online banking access application. He plans to poll customers in a systematic manner (1 out of every 5), 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 the following questions.

1. What sample size should be chosen?
2. Will this type of sampling method be effective in estimating the population?
3. What measures of central tendency, dispersion and graphics would be effective to tell a story with the data?
4. How to ask the right questions to gather the proper loyalty data and then how should the probability of customer attrition be calculated?

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

## 1.1 Descriptive Statistics

Descriptive statistics represent the foundation of any statistical analysis. When one has a data file and wants to begin to understand the data the first step is to describe the data. Descriptive statistics are the tool used to accomplish this objective. Human beings cannot make sense of more than a few rows of data organized in a column from a spreadsheet without the help of descriptive statistics. 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 the value are to one another. Descriptive Statistics are often reported in conjunction with one another as any single number or univariate statistic is usually not sufficient in fully explaining the characteristics of the variables in the dataset. For example, its misleading to report the mean as an estimator of the population if the data are highly skewed. Moreover, there are some phenomena which are bimodal and therefore should 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).

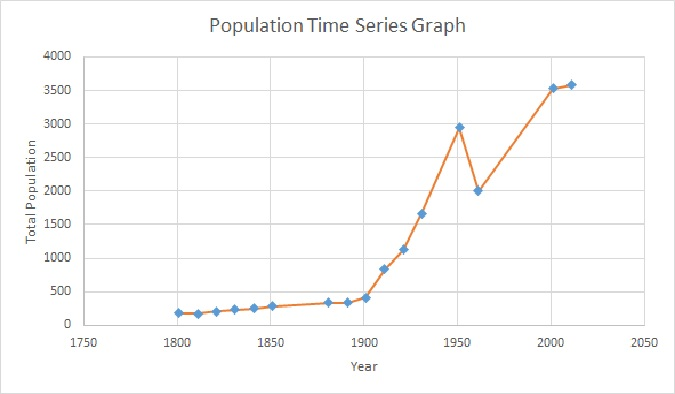
Descriptive statistics is often confused with being a lower form of data analysis on a maturity curve representing the competitive advantages of analytics. This is a misnomer as descriptive statistics are needed to support many higher-level analyses as will be presented throughout the course.

### Data Terminology

The first discussion we need to have when beginning a study about statistics is an overview of data terminology. When discussing data, we need to define avariable which is a characteristic of a subject or object which 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 collection of observations from the variables, while a data set is all of the values of one of more variables collected for research purposes. Many are often first exposed to variables, data, and datasets via Excel spreadsheets.

Next, we want to cover the types of data. Categorical data also known as qualitative data are data which are described in words rather than numerically. An example of a categorical variable would be car color. Numeric data, on the other hand, come from counting measurements or a type of mathematical operation. Examples of numeric variables include income and time of day. The two major types of numeric data include discrete variables and continuous variables. Discrete variables are those variables which can be counted, such as the number of customers entering a store is an example of a discrete variable. Continuous data represents those variables which cannot be counted; things like weight and time are examples of continuous variables, as these can both be measured ad infinitum. Another type of data worth mentioning, especially in relation to business analytics, is time-series data. Time series data are data which are measured consecutively over a period of time. Below is a graph representing time-series data.

**Time Series Graph**



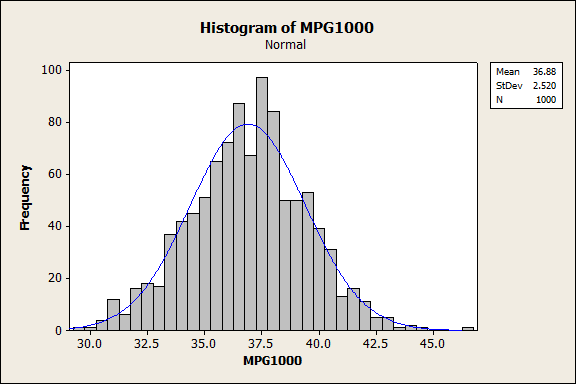
### Levels of Measurement

The levels of measurement for data are important in choosing the right statistical techniques to incorporate. The four levels of measurement for data are nominal, ordinal, interval, and ratio. Nominal data derived from the Latin “nomen or name” identifies a category (Doane, 2016). Categorical data like hair color would be measured at the nominal level. Next, we have the ordinal level of measurement. Ordinal data are ranked data, but the difference between observations is not meaningful outside of the ranks. Credit ratings of companies are an example of data at the ordinal level of measurement. After that is 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 as they are not fully known. Fahrenheit or Celsius measures of temperature are examples of measurements or variables at the interval level of measurement. 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 measured and makes the use of ratios possible with the data. The temperature measure Kelvin is an example of a variable at the ration level of measurement (as opposed to the previously mentioned Fahrenheit or Celsius measures) as Kelvin is measuring the kinetic energy of small particles using the law of thermodynamics and therefore has a natural zero unlike the other measures. Ratio and interval levels of measurement are most suitable for statistical techniques. 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 someone or a corporate entity below zero. Nonetheless, there would still be a point at which credit is no longer obtainable thus making the natural zero point (Ahmad, 2016; Doane, 2016).

### Measures of Center

Now that we have covered an overview of the characteristics of the data, we will now be discussing 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 below shows this tendency for a typical variable.

**Histogram**



Before we delve into the measures of center some basic statistical notation is necessary. Please review a partial list of statistical notations below (Byjus Classes, 2021).

|  |  |
| --- | --- |
| *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 |
| *s* 2 = | sample variance |
| X ~ = | distribution of X |
| *zx* = | standard score |

### Mean

**Mean**

If the data are normally distributed, the mean should always be chosen as the measure of center because it can be used later in more powerful statistical techniques.

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

(1.1)

We can also calculate the mean using the Excel function =AVERAGE (data), where data is a column variable containing the actual data. It should be noted that the mean is only suited for variables at the interval or ratio levels of measurement (Microsoft, n.d.).

The arithmetic mean aka “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 is subject to not as accurate as 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 sorted data variable. If the data have an uneven number of observations, then the median is calculated by taking the average of the middle two observations. The median should be used in place of the mean when there are extreme values. However, it should be noted that the median lacks some of the useful mathematical properties of the mean making it less versatile (Doane, 2016).

### Mode

The moderepresents the most frequent value in the data. If the data are tending to be centrally distributed then the mean, median and mode can be similar. However, the mode can also not be present at all in a dataset and a dataset may also have several frequent variables making it bimodal. To identify the mode one has to observe the frequency of each value of a variable using the Excel function = MODE.SNGL (data variable) (Microsoft, n.d.). The mode is frequently used to describe central tendency in a 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 frequently used. Please note similar to above that in a symmetrical distribution where the observations cluster around the center (normally distributed) the mean, median and the mode can be very similar.

### Measures of Dispersion

The shape of a distribution can be observed through the use of a histogram as displayed in an earlier part of this section. One can also compare the mean and median If the data are skewed right it signifies that the mean exceeds the median. On the other hand, if the data are skewed left, this is an indication that the mean is below the median. When the mean and media and even the mode are similar then the distribution is said to be symmetric. The difference of the the data from 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 rangewhich is the highest value of the data subtracted from the lowest 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 calculate using the function = MAX (data variable) - MIN (data variable) (Microsoft, n.d.). The most popular and versatile measure of dispersion is the standard deviation(Ahmad, 2016). The standard deviation measures the average distance between the 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 in order 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 use by researchers there are also other important ones including measures of 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.Your readings will provide useful information in these and other descriptive statistics.

### Descriptive Statistics Summary

Descriptive statistics are important tools which should be a part of any data analysis. Anyone tasked with the problem of analyzing a dataset should incorporate descriptive statistics as an automatic first step. Descriptive statistics allow researchers to evaluate a dataset and help them to choose the most appropriate more advanced statistical technique to incorporate in future analyses. One will be or are already used to seeing a great deal of descriptive statistics in any standard business presentation.

### Self-Check Questions

1. Why must we use descriptive statistics to describe the data we collected?

* *They allow researchers to describe variables in a dataset.*
* They describe where the data originated from.
* They indicate who collected the data.

1. Why are measures of dispersion important descriptive statistics?

* *They measure the extent of variation in the data.*
* They measure the value of the data.
* They are the only measures needed to describe the data.

1. Which of the following is an example of numeric data?

* *Time*
* Gender
* Marital status

1. Which of the following is an example of a measure of center?

* Standard deviation
* Range
* *Mode*

## 1.2 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.

With **inferential statistics** we are primarily concerned with “inferring” information on a population from a sample. In many cases it is either impractical or impossible to gather information from a population. Two common cases are political 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 provided appropriate probability sampling methods are conducted.This represents a huge cost savings in terms of sampling rather than taking a sample of the population. Moreover, if you were tasked with understand how many salmon there were currently in the Columbia River this would obviously be an impossible task. However, with the correct sampling planning and execution this could be fairly straightforward. In some cases, like in conducting a census of the entire population of a country, which is often mandated by law, a survey can even be more accurate since 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 under inferential statistics include estimating parameters, testing hypotheses and regression and trend analysis. There are also many aspects of quality control which make use of inferential statistics. We will be covering many inferential statistics in later sections.

**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.

In order to perform inferential statistics, it's important to sample the data from a population in a proper manner. This is to ensure that the inferential statistics one would calculate are unbiased estimators if the population. Random sampling methods include simple random sampling 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 has a chance of being selected which is greater than zero. Types of **probability sampling** in addition to simple random sampling include systematic sampling, cluster sampling and stratified sampling. Since these methods all give elements of the population, a chance of being selected these are considered to be the most desirable sampling methods for estimating a population using inferential statistics (Knapp, 1980). Other sampling methods where all members or elements of a population do not have a chance of being selected include convenience sampling, volunteer sampling and snowball sampling. While methods like convenience sampling which draw samples based on convenient segments of a population tend to be popular in practice, they may severely bias the estimates of the population using inferential statistics (Doane, 2016; Knapp, 1980).

Once a sampling method has been chosen the next step is to collect data from a sample. One of the primary methods of collecting data from a sample is through a questionnaire. 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 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 statistics derived from a sample. A statistic is a measure that is derived from a sample. A parameter is a measure describing the entire population. All samples have random error and therefore the difference between a parameter is 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 take sampling error into account.

A point estimate is a single value estimate of a parameter. A mean would be 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 one if the most popular interval estimates which 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.

With a 95% confidence interval it means that with an infinite number of samples the interval will contain the true population parameter 95 out of 100 times. The calculation for a confidence interval can be found in equation (1.4) below:

Hypothesis testing, correlation analysis and regression analysis are addition inferential statistics which will be covered in later sections.

(1.4)

where:

CI =Confidence Interval

top enclose x = sample mean

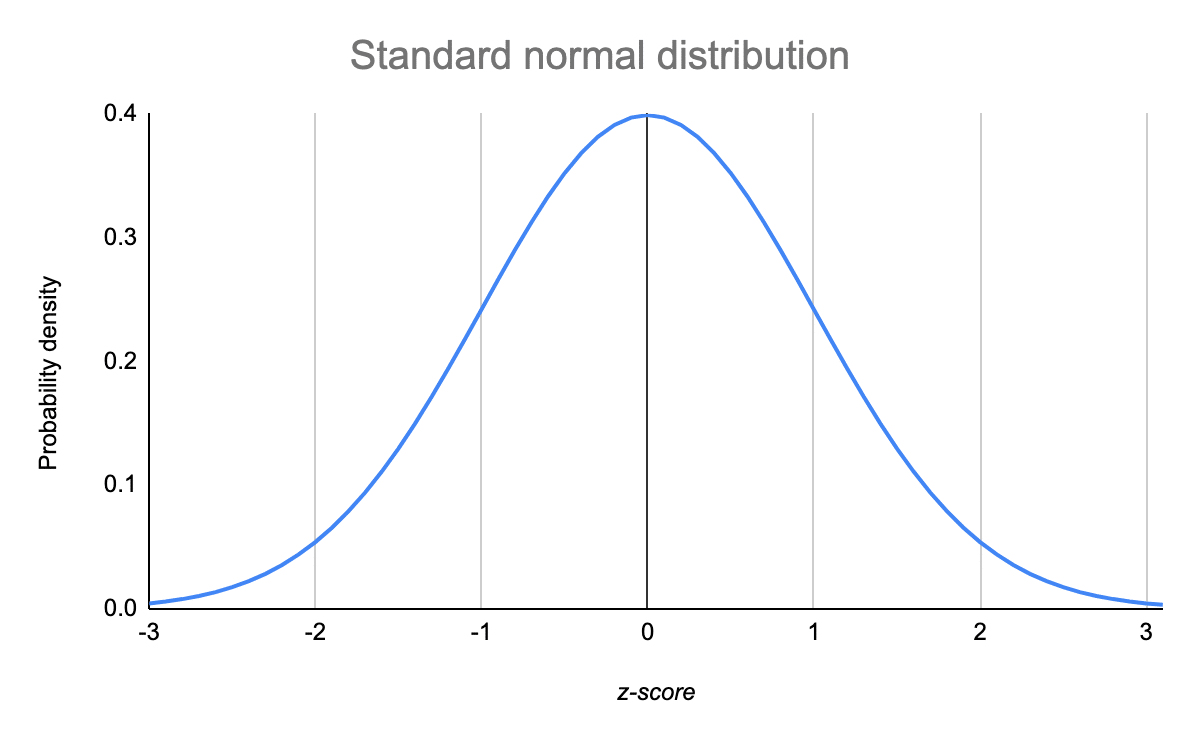
z = confidence level value

s = sample standard deviation

n = sample size

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 unit normal distribution (Doane, 2016; Knapp, 1980). The figure below depicts the standard normal distribution.

**Standard Normal 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).

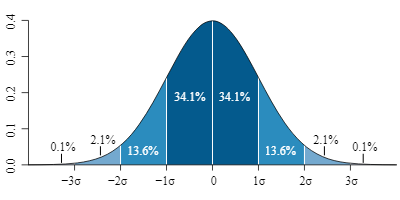
### Z-Calculation

The calculation for the z z-score is depicted below, where x represents the score, mu μ represents the mean, and sigma σ represents the standard deviation. shown in the equation (1.5) below.

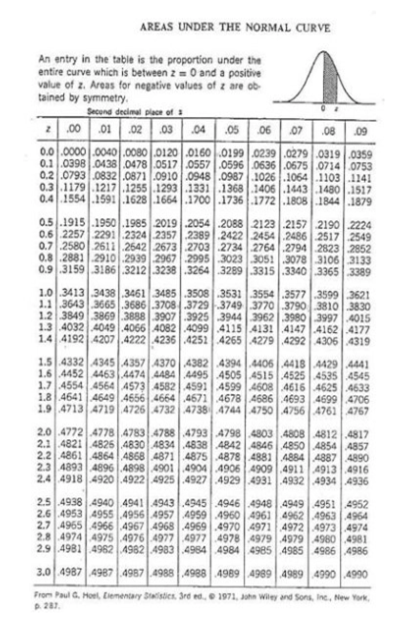
(1.5)

Once you calculate a z-score one can look it up in a z-table which is depicted partially below. For access to the full table see (Farber & Larson, 2017). The z-table represents the probabilities under the standard normal distribution, which enables researchers to calculate the probability of randomly obtaining a score from the sample distribution. For example, there is a 95.45% 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).

**Standard Normal Distribution Probabilities**

.

**Areas Under the Normal Curve Example**



Inferential statistics are a powerful set of tools which 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 a very efficient and sustainable manner and is an important tool to study and continually use in a business environment.

### Self-Check Questions

1. What does a z-score tell us?

* *How many standard deviations from the mean each value lies.*
* They measure the value of the data.
* They are the only measures needed to describe the data.

1. Please complete the following sentence:

The standard normal distribution has a mean equal to *0* and a standard deviation equal to *1*.

1. Please complete the following sentence:

In a simple random sample, each unit in the population has an *equal* chance of being selected.

1. Please complete the following sentence:

In a standard normal distribution, there is a *68.2%* probability of randomly selecting a score between -1 and +1 standard deviations from the mean.

## 1.3 Probability Calculation

Archeological digs have discovered that ancient societies have played some sort of probability games with primitive dice-like objects made from animal bones. However, it wasn’t until people created organized gambling that the serious study of probability was commenced. There are still many outstanding questions on what true probability means. Classical physics would state that true probability is due to the fact that we have unknown inputs. Albert Einstein famously commented that “God doesn’t play dice”. However, quantum physics which dominate our lives in 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 the equal odds of something occurring. For example, rolling a die deemed to be fair would have an equal likelihood of getting 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. Finally, choosing any one card from a standard deck of cards produced a probability of 1/52 (Di Paola et al., 2018; Doane, 2016).

### Formulas for Classical Probability

Classical or theoretical probability is the expected probability for a simple event happening (i.e., rolling dice). This is represented by the number of times the event can occur (1 roll of the dice) divided by the number of possible events (6 sides).

P(A) means “probability of event A” (event A is whatever event you are looking for, like rolling a die and getting a 6). “f” is the frequency, or number of possible times the event could happen, and N is the number of times the event could happen.

### Examples

The odds of rolling a 3 on a fair die are one out of 6, or 1/6. That’s one possible outcome divided by the number of possible outcomes (6). The odds of winning the Powerball lottery are 1/292.2 million. The “1” is the number of times the event can happen (meaning you win), 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.

### 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 the number of possible outcomes, but instead seeks to calculate the actual probability of events occurring in an experiment (Doane, 2016). An interesting example of this was a statistician soldier from the US who was a prisoner of war during the Vietnam Conflict recorded the results from hundreds of thousands of repeated coin flips done during his time in the prison cell and found the **empirical probability** was slightly biased toward heads.Classical 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 can be given as in equation (1.6).

**Empirical probability**

Sometimes, even the most widely accepted classical probabilities (e.g., coin toss probability) can be upended by empirical probabilities.

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 of the event attempts to forecast what will happen on the basis of total no. of outcomes possible. We expect both the experimental and theoretical probabilities to converge as the number of trials in an experiment increase (Di Paola, et al., 2018; Doane, 2016).

### Subjective Probability

**Subjective probability**

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

Subjective probability is used when there is the absence of a random repeatable experiment. Subjective probability uses an informed judgment to predict the likelihood of an event. Interestingly, we use **subjective probabilities** in our everyday lives. Some real-life examples where humans may use subjective probability are

* the outcome of a job interview,
* getting an accepted bid on a house,
* the risk of not wearing a seat belt, and
* the odds of seeing a ghost when walking through a graveyard at night.

### The Seven Basic Rules of Probability

The basic rules of probability are both logical and intuitive. Sometimes, probability problems can get 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 one will have to apply both their logic and counting skills (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 0 and 1. The probability calculation is incorrect if it's less than zero or greater than 1 (Di Paola et al., 2018; Doane, 2016). Equation (1.7) depicts probability rule one.

For an event A, 0 ≤ P(A) ≤ 1.

(1.7)

#### Probability rule two

The first basic rule of probability is that all probabilities are represented by a number between 0 and 1. The probability calculation is incorrect if it's less than zero or greater than 1 Di Paola, et al., (2018) (Doane, 2016). This is depicted in the following equation:

For an event A, 0 ≤ P(A) ≤ 1

(1.8)

#### Probability rule three

The second basic rule of probability is that the combined outcomes of all probabilities must be equal to 1 (Di Paola et al., 2018; Doane, 2016). Equation (1.9) depicts probability rule three.

Sum of all outcomes = 1

(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 even will happen Di Paola, et al., (2018) (Doane, 2016). Equation (1.10) depicts the fourth probability rule. 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 the probability of two events is the sum of the probability that will happen minus the probability that both will happen Di Paola, et al., (2018) (Doane, 2016). Equation (1.11) depicts probability rule five, the general addition rule.

P(A or B) = P(A) + P(B) – P(A and B)

(1.11)

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

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

In this section 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 the very important concepts of independent events and conditional probability. Two events A and B are stated to be independent if the fact that one event has occurred does not affect the probability that the other event will occur. However, whether or not one event occurs does affect the probability that the other event will occur, then the two events are stated to be 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 independent events. This rule states that of 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). Equation (1.12) depicts probability rule six. If A and B are two INDEPENDENT events, then

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

(1.12)

#### Probability rule seven

The final and seventh 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 of probability is known as the conditional rule of probability. This rule states that the conditional probability of event B, given event A is the probability of both events A and A occurring divided by the probability of event A occurring (Di Paola et al., 2018; Doane, 2016). Equation (1.13) depicts probability rule seven.The conditional probability of event B, given event A, is

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

(1.13)

### Probability Summary

So now we covered the seven basic rules of probability. Understanding probability is critical to the understanding of the inferential statistics this course will be covering in upcoming sections. Please refer to your readings for more details on these concepts. One should be advised to always use their best logic when solving probability problems and to apply these rules always in a systematic manner. Sometimes, our minds are seeking a simple solution and can easily overlook some of the nuances of a probability problem so please take the time to follow a logical train of thought whenever you encounter a probability problem.

### Self-Check Questions

1. Which of the following is an example of an empirical probability?

* *The probability of a sunny day in Phoenix AZ is 98 percent based on previous weather data.*
* The chance of rolling a six in a dice throw is 1/6.
* The chance of landing on heads in a coin flip is 50/50.

1. Please complete the following sentence:

Classical or theoretical probability is represented by the number of times the *event* can occur divided by the number of possible *trials*.

1. Which of the following is an example of a subjective probability?

* *The probability of winning the business contract is estimated by the sales manager to be about 50/50.*
* The chance of rolling a three in a dice throw is 1/6.
* The chance of drawing an ace of spades from a deck of cards is 1/52.

1. Please complete the following sentence:

The second basic rule of probability is that the combined outcomes of all probabilities must be equal to 1.

Summary

Unit One exposed students to some of the most basic and foundational statistical principles including descriptive statistics, inferential statistics and basic probability analysis. These are foundational principles which could be used as the building block for the more advanced studies in multivariate statistical tools and concepts . As such students should make sure they have a good grounding in each of the topics we covered and should bookmark some of these key concepts as they may need them as they move ahead in acquiring statistical knowledge. If students can have a good grounding in each of these topics then they will be fully prepared to learn more advanced techniques and will be comfortable in digging into any dataset they receive as a part of their work in order to extract insights necessary to make the right business decisions.

# Unit 2 – Bivariate Analysis

### Study Goals

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

… recognize how 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 United Kingdom grocery store chain Lesko want to know if its coupons offering a 5% discount 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, so Catherine’s concern is that these are the wrong target customers, since they are most likely to be coming into the store to shop anyway. Since all of Lesko’s current loyalty subscribers are receiving these coupons, Catherine ponders 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 very questions Catherine knows she will need to address before she receives the desired information.

**Crosstabulations**

These can also be used to validate statistical models.

This unit is an introduction tobivariate analysis. Bivariate analysis represents a series of statistical analysis procedures on two variables. This is in contrast with univariate statistics (for a single variable) and multivariate statistics (for three or more variables). Bivariate analysis focuses on situations with one dependent and one independent variable. The denotation of these variables are X and Y. With bivariate analyses we analyze if there are changes that occur between these two variables and how much.

**Market research**

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

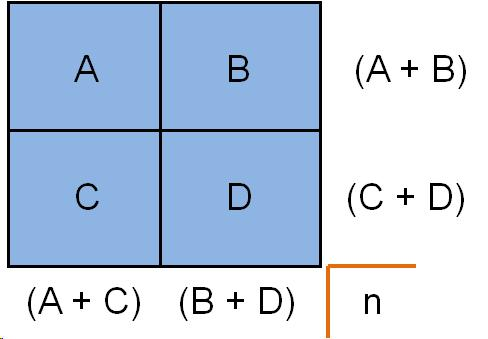
## 2.1 Crosstabulations

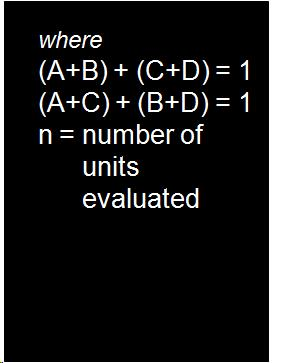
**Crosstabulations**, also known as crosstabs, are the bread and butter of packaged goods firms. These firms use **market research** firms to conduct field research and do extensive surveys of consumers and current customers and use crosstabulations to segment their customers and extract insights on the use of their products. Examples include a previous use of the product in the last 30 days by various age brackets. These two variables can be arranged in a table for researchers to then be able to ascertain if certain age groups are using the product more often. This is just one example of the multiple possibilities with crosstabulation analysis (Doane, 2016).

A crosstab for two variables is called a 2-way contingency table and forms a table with a rectangular format with rows represented by R categories of the X variable and columns represented by C categories of a Y variable. Cells are represented by each intersection, and these represent the possible outcomes. The frequency 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 below figure is an example of a 2 x 2 table.

**2 x 2 Crosstabulation**





A **crosstabulation** contains three probability distributions one needs to be concerned with: joint, marginal, and conditional(Doane, 2016). The joint probability distribution in a crosstab describes the proportion of the subjects jointly classified by a category of X and a category of Y. Dividing the cells of the crosstab by the total provides the joint distribution. The sum of the joint distribution is 1 as indicated in the crosstabulation graph above (Doane, 2016).

**Crosstabulation**

The statistical significance of a crosstabulation is usually measured using non-paramentric statistical techniques, including the chi-square analysis.

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 1 as indicated in the crosstabulation graph above (Doane, 2016).

The distribution of one variable given the levels of the other variable is described through 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 conditional distribution is 1 as indicated the crosstabulation graph above (Doane, 2016).

There are many other rules with crosstabulation tables which can be further researched in the readings including 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).

### Self-Check Questions

1. Which of the following is an example of probability distribution of interest in a crosstabulation table?

* A priori probability distribution
* Judgmental probability distribution
* *Marginal probability distribution*

1. Please complete the following sentence:

Another name for crosstabulation table or crosstab is a contingency table.

3. A contingency table having R rows and C columns is referred to as what?

* 3x2 table.
* BxD table
* *RxC table*
* Summary table

4. Please complete the following sentence:

The distribution of one variable given the levels of the other variable is described through the *conditional* probability distributions.

## 2.2 Mean Comparison Test

How do we know if the difference between two means/medians is due to a significant treatment effect of is just due to random variation? We already know from the study of inferential statistics that the sample mean can vary from the population due to random sampling error. Therefore, when we see a difference between two sample means/medians how do we really know if this difference is due to random error or some treatment effect or business activity that resulted in a significant difference? Our study of mean/median comparison tests will enable one to detect this difference. We are going to go into details on hypothesis testing in later sections, but it is important in this section on bivariate analysis to introduce the concepts and terminology of comparing two means or medians. As we will cover the t-test in a later section we will first introduce a **nonparametric** test to compare medians in this section to demonstrate the concept. This test is known as the Wilcoxon signed-rank test. Non-parametric tests are simpler tests than parametric tests and do not rely on the as many assumptions on the distribution of the data than parametric tests. Non-parametric literally refers to data not having a normal distribution. And a non-parametric test therefore does not depend on the data being normally distributed. The Wilcoxon-signed ranks test is used with dependent samples with these being the most interesting in studying bivariate analysis since we are mainly focused on the relationship between two variables at this stage. The Wilcoxon signed rank test actually 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 for as we learned in descriptive statistics the median is less susceptible to outliers in the data (Fogarty, 2017).

**Nonparametric**

Statistical techniques that are nonparametric do not require the same assumptions of the distributions of the data as parametric techniques. However, there is a tradeoff in power of the statistical technique.

An example of two dependent means would be for example the results of a pretest for knowledge in corporate ethics compared to a posttest taken after employees received online ethics training. We can conclude that the two samples are dependent or related to one another since the same employees are taking the test before and after the training.

The Wilcoxon signed rank test (aka the Wilcoxon signed rank sum tests) was developed by Frank Wilcoxon while doing chemical compound research at American Cyanamid Labs in Stamford, Connecticut for such products as Old Spice, Combat Insecticides and Pine Sol Cleanser (Fogarty, 2017).

There are two slightly different version 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 previous case study we would want to know whether our training had any affect (Doane, 2016).

Remember for data which is normally distributed assuming a large enough sample the mean and the median are equal. The null and alternate hypothesis will covered in later sections but for the purposes of the Wilcoxon test, the null **hypothesis** is that the medians of the two samples are equal. This indicates there is no treatment effect (Doane, 2016).

**Hypothesis**

Hypothesis tests are critical in determining whether the differences between medians are due to a treatment or program rather than a random sampling error.

The research question we are testing from the example above is that the difference between the pretest and posttest mean is statistically significant due to the corporate ethics training.

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 available in your readings for categorical dependent variables or use 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 pretest 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 by *T* the unsigned value of the smaller sum of ranks; *n* is sample size (number of ranks), *μ* = mean, σ2 = 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 from the readings for a [one-tailed test](https://www.sciencedirect.com/topics/medicine-and-dentistry/one-tailed-test) as appropriate.

If the p-value is very low (i.e., < .05) then this indicates that is would be a very rare occurrence that the difference between means occurred 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 vaccines and other lifesaving medicines, have been discovered via hypothesis testing.

Self-Check Questions

1. Which of the following is an example of why a statistical test is needed to compare two means or medians?

* *It determines the effect over random sampling error.*
* Because the data are uncertain.
* It increases the probabilities related to the data.

2. For running a Wilcoxon signed ranks test, the dependent variable must at least be categorizable as which of the following?

* Categorical
* Continuous
* Normal
* *Ordinal*

3. Please complete the following sentence:

Means calculated on a pretest and posttest scores applied to the same individuals are known as *dependent* means.

4. What does a very low p-value for the Wilcoxon signed ranks test suggest?

* The difference in means was due to the treatment.
* There is a problem with the test.
* *The random sampling area is probably the reason for the differences in means.*
* The wrong test has been chosen.

## 2.3 Correlations

When one wants to determine the direction of the linear relationship between the two variables the covariancecan be used. The values of covariance can be any number between negative and positive infinity. One can determine the value of covariance between two variables by summing the product of the differences from the means of the variables using the following equation (2.1):

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 but has some useful features such as remaining unaffected by the change in scale, dimensions and location, and can also be used to compare two pairs of variables across different domains (Doane, 2016).

The correlation is determined by normalizing the covariance. This is accomplished by dividing it with the product of the standard deviations of the two variables using the equation (2.2) below (Doane, 2016).

(2.2)

where:

* Cov is the covariance
* σx is the standard deviation of X
* σy is the standard deviation of Y

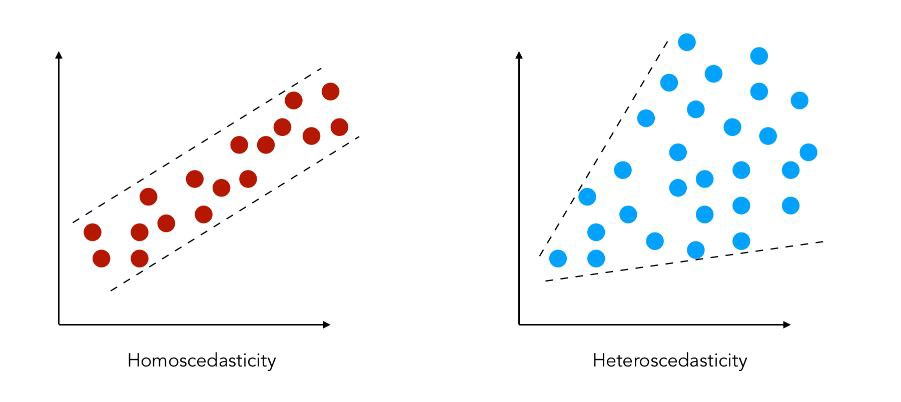
The correlation coefficient measures the strength and direction of the relationship between two variables. It is a measure of the degree of “linear” relationship between two variables. The most common is known as Pearson Product Moment Correlation, noted in the literature as “r” (Doane, 2016).

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 published in 1896 titled “Mathematical Contributions to the Theory of Evolution III. Regression, Heredity, and Panmixia” (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 correlation coefficient indicates the percentage of variation in one of the variables due to knowing the other variable. On must be made aware of the fact the correlation coefficient Correlation only measures the straight-line or “linear” relationship between variables. Correlation does not measure non-linear relationships very accurately. There are other assumptions with correlation including homoscedastity. The homoscedastity assumption is that the variance is the same for all values of variables in the y axis of a scatterplot. Assuming homoscedasticity is accepting that variance is fixed throughout a distribution. When there are serious violations of the homoscedastity assumption it underestimated the Pearson product moment correlation coefficient (Doane, 2016).

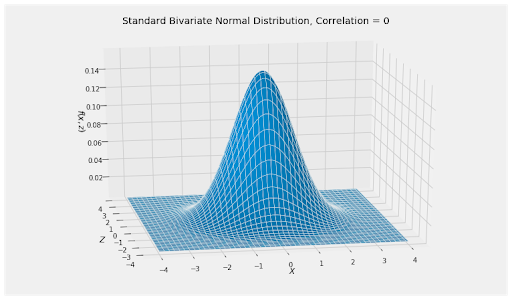
The heteroscedasticity assumption with correlation results from a nonnormality of one of the variables (as shown in the figure below), an indirect relationship between variables, or to the effect of a data transformation. It is important to note that heteroscedasticity does not invalidate a correlation analysis, but instead 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 is rectified through transformations of the variable (Doane, 2016).

**Homoscedasticity and Heteroscedasticity**



Correlation also assumes that each pair of variables have a bivariate normal distribution. 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.

**Standard Bivariate Normal Distribution**



Additional issues researchers need to be aware of with correlation analysis are non-linear 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 and 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 to not get the wrong interpretation. Finally, there can be truncated ranges of the variables 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” variable, X) (Doane, 2016).

2. Multiple correlation (between one “dependent” variable, Y, and many “explanatory” variables (Doane, 2016).

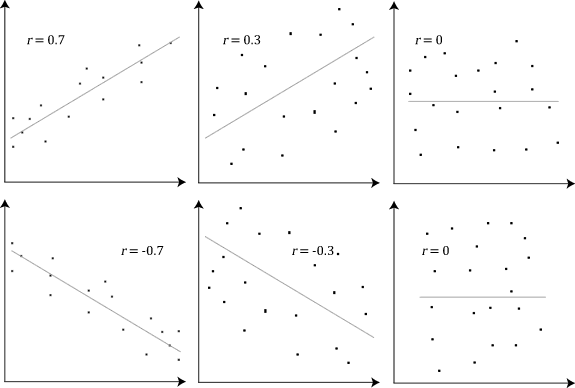
3. Canonical correlation (between many “dependent” variables and many “explanatory” variables) (Doane, 2016)

Simple correlation as stated above is calculated between two variables. Multiple correlation is correlation that is computed between one variable and two or more variables (e.g., the relationship between the body mass index of an individual as a function of height, age, and average daily caloric intake). So, it involves using many variables (Doane, 2016).

It is important to note that correlation 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 is associated with increases in the Y variable (and vice versa). A negative simple correlation tells us that increases in the X variable is associated with decreases in the Y variable (and vice versa) (Akoglu, 2018).

This “scatterplots” below show, graphically, 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.

**Correlation Coefficient Examples**



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.

**Spurious correlations**

These can be prevented by always having a strong hypothesis as to why the correlation is occurring. Without a hypothesis, it is very easy to find correlations in the data that don’t really exist.

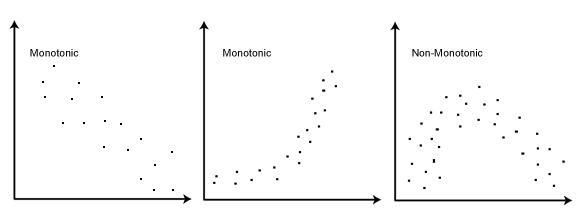
For instance, with four variables, there would be four different simple correlations computed; the matrix (table) would consist of these four correlations in the off-diagonal elements, and the number one (1) in the main diagonal of the table. **Spurious correlations** can also occur when two variables seem to be correlated (numerically) but are not actually. Often, their correlation is really driven by a third, hidden variable.

Correlation and regression analysis go hand in hand and sometimes one is interchanged for another depending upon the application. While correlation measures the strength and direction of relationship, it does not give the actual linear relationship. Regression analysis will yield the equation of the straight line (the blue line in the plots above) going thru 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 mathematical 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 which can be efficient when the two variables have mixed levels of measurement. Usually, in statistics, we measure three types of correlations in addition to the Pearson correlation. These include 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 variables are not linearly related (i.e., an exponential relationship), one or both of your variables are ordinal variables, or if there are outliers in the data. If either 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 correlation 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 the ρ will always be a value between –1 and 1 as Spearman’s correlation is equivalent to calculating and interpreting the Pearson Product Moment correlation coefficient. Finally, it should be noted that the Spearman’s correlation assumes that the relationship between the two variables being evaluated is monotonic. In other words, and 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.

**Monotonic vs. Non-Monotonic Relationships**



Another non-parametric version of the Pearson correlation is known as Kendall’s Tau Coefficient. The Kendall’s Tau also known as the Kendall rank coefficient uses ranks of the data in order to assess statistical associations. Both the Kendall’s Tau and previously discussed Spearman’s Rank coefficient are good alternatives to the Pearson correlation when one or more assumptions of the Pearson correlation could not be met. Kendall’s Tau is often used as an alternative method to the Spearman correlation when the data has many tied ranks and or when 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 observes whether there is a pattern of concordance or discordance between the pairs determines the strength of association based on the patter on concordance and discordance between the pairs (Akoglu, 2018).

* Concordant**:** Pairs are ordered in the same way (consistent). A pair of observations is considered concordant if (x2 — x1) and (y2 — y1) have the same sign.
* Discordant**:** Pairs are ordered differently (inconsistent). 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 relationship between variables (Akoglu, 2018).

Correlation analysis in its many forms represents 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.

### Self-Check Questions

1. Which of the following is an assumption of the Pearson product moment correlation?

* *The X and Y variables for a bivariate normal distribution*
* The relationship between the two variables is nonlinear
* At least one of the variables has a mean of greater than 1

1. Please complete the following sentence:

Some phenomena have a cause and effect. However, correlation does not necessarily imply *causation*.

1. Please complete the following sentence:

Correlation and *regression* analysis go hand in hand, and sometimes one is interchanged for another depending upon the application.

4. When there is no relationship between two variables the correlation coefficient must be:

* > 1
* = 1
* < 0
* *= 0*

Summary

In this unit, bivariate statistics were covered, which gave a first look at how to measure and describe the relationship between two 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. In the modern world of big data, bivariate statistics quantifying relationships between variables are important to be able to understand factors which may be related to phenomena we observe in business and in our personal lives. However, we must be still cautious when discovering these relationships as spurious correlations can occur and we must always remember that correlation does not imply causation. However, we should not be too cautious as there is still no causation without correlation.

2.1

The use of crosstabs represents one of the most basic and simple ways to quantify a relationship between two discrete or more variables. Crosstabs are used in many business analysis projects not only as a primary outcome of a business research study but also as a preparation phase in order to understand what features one need to add to more sophisticated models.

2.2

# 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 German-based company Riedling is a biotechnology firm specializing in creating better healthcare solutions for an aging population. The company has recently been gathering data around its solutions in order to be able 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 were in market it wanted to gain a longer-term view. The data gathered includes all of the physical information from the patients of the trial which includes height, weight, age, gender, marital status just to name a few. Also added were health measurements like blood pressure, cholesterol, and blood sugar levels. Data was 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 analysis on the data and set out a first task of classifying all of the variables as discrete or continuous and calculate the means, standard deviations and other descriptive statistics. This he learned from his training in descriptive statistics. However, this time he wanted to go a step further and produce histograms for all the data knowing that if his data was approximately normal, he could run techniques like correlation analysis and general linear models. He immediately noticed that variables like height, weight, blood pressure were normally distributed and other measurements like health engagement were not normal. Harvey began to create a new spreadsheet of all the variables which were normally distributed versus those which 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 or is just the lack of information about the universe. The famous physicist Albert Einstein was once quoted as saying “God does not play dice with the universe”. However, scientists trained under the quantum physics paradigm would argue that there is true randomness in the universe and have quantum experiments to lend evidence to this conjecture. Whatever the orientation toward this big argument 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 variables. Typically, **random variables** are defined as those variables derived from experiments, however, they can also be derived from processes happening on a continuous basis around us. Random variables are critical components to statistical analysis and your knowledge of them as students is absolutely required to gain a full understanding of probability distributions (McEvoy, 2018).

**Random variables**

Unlike algebraic variables, 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 statistical calculations. One of the most typical random variables is the coin toss. If we toss the coin one hundred times and each time we record the outcome in a journal the recorded outcome represents a random variable. Random variables are classified into two types which are discrete and continuous. Discrete random variables like the number of students in a classroom can be determined through the process of counting. Continuous random variables on the other hand can be determined only through some type of measurement. An example of a continuous random variable is time which can be broken down, for example into centuries, decades, years, days, hours, minutes, seconds, or miliseconds (McEvoy, 2018).

Random variables emerging from experiments which can be either continuous or discrete are important in statistics because these variables all follow distributions which have properties useful to conducting statistical analysis. Discrete random variables which include examples such as the number of cats in a litter, the number of people attending a basketball game, the number of students in attendance in class and the number of people in line at the department of motor vehicles on a given time and day. Continuous random variables on the other hand cannot be counted and they can take infinitely many values. In other words, continuous random variables have infinite precision. Continuous random variables can only be determined through some type of measurement. Continuous random variables include distance, temperature, pressure, height, weight, mass, density, temperature, and volume. It is also useful to note that there are no absolutes and discrete random variables can be partially continuous and moreover continuous random variables can be partly continuous. These variables are referred to as mixed type variables.

Self-Check Questions

1. In statistical terms, how is randomness measured in our lives?

* *Random variables*
* Monte Carlo simulations
* Coin tosses
* Random number generator

1. Please complete the following sentence:

Random variables are typically those derived as a result of random *experiments*.

1. Please complete the following sentence:

Discrete random variables can be determined through the process of *counting*.

1. Continuous random variables can only be determined by which of the following methods?

* Counting
* *Measurement*
* Hypothetical distributions
* Best judgement

**Discrete probability distribution**

A discrete probability distribution represents the probability of the occurrence of each value of a discrete random variable. With a discrete probability distribution, each possible value of the discrete random variable can be associated with a probability that is greater than zero. Thus, a table or histogram is often used to present a discrete probability distribution.

## 3.2 Discrete Probability Distributions

Discrete random variables have a distribution which is identified as a **discrete probability distribution** (Viti, Terzi, & Bertolaccini, 2015; McEvoy, 2018). The discrete probability distribution of a discrete random variable is a catalog of probabilities associated with each of its possible values. It is also sometimes called the probability mass function or just probability function. Some of the most popular discrete probability include the binomial probability distribution, hypergeometric probability distribution, multinomial probability distribution, negative binomial distribution, and the Poisson probability distribution (Viti, Terzi, & Bertolaccini, 2015; McEvoy, 2018).

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. . Moreover, the sum of all the probabilities is Shape

Description automatically generated with medium confidence (Viti et al., 2015; McEvoy, 2018). Discrete probability distributions can be visualized via a histogram. Below is a histogram 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 first will 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 each time record and plot the probability of each possible outcome, 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).

Discrete Probability Distribution

Chart, histogram

Description automatically generated

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. . 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 distribution plot. This histogram is obtained by recording the probability of each possible outcome if a discrete random variable over many trials and then plotting the probability on the y-axis and the outcome of the discrete variable on the x-axis. There are many different types of discrete random variables. Some popular ones include the binomial probability distribution and the hypergeometric probability distribution (Viti et al., 2015; McEvoy, 2018).

Self-Check Questions

1. A discrete probability distribution is the probability distribution of this type of variable?

* Continuous random variable.
* *Discrete random variable*.
* Missing variable.
* Independent regression variable

1. Please complete the following sentence:

The probability of each value of a discrete random variable ranges between 0 and *1*.

1. Please complete the following sentence:

A discrete probability distribution is sometimes referred to as a probability *mass* function.

1. Which of the following random variables would have a discrete probability distribution?

* Time of day
* *Results of a coin toss*
* Temperature
* Distance

3.3 Continuous Probability Distributions

**Continuous probability distribution**

These are typically plotted as a curve, the area of which represents the cumulative probability. They typically represent the probability by utilizing specific formulas, geometry, technology, or probability tables.

If the random variable is continuous than 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). A discrete probability distribution has a range of values that are countable. For example, if Hallmark wanted to create birthday cards for all possible ages, they would range from 0 to 122 with 122 representing the oldest person that ever lived. A continuous distribution of random variable on the other hand has a [range of values](https://www.calculushowto.com/types-of-functions/domain-and-range-of-a-function/#definition) that are [infinite](https://www.calculushowto.com/is-infinity-a-number/), and therefore, are 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 past10 years has always been between 34 to 44 degrees centigrade. The temperature can take any value between the ranges 34 to 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∘ to 39.5.5∘. Any observation which is taken falls in the interval (Crooks, 2019).

**Probability density function**

Since many of these distributions are not closed-form solutions, integral calculus is used to find the area under the curve as represented by the probability density function.

In addition to the differences between discrete and continuous probability distributions stated above the use of a formula to express a continuous probability distribution is another difference (Crooks, 2019). While discrete probability distributions are often described with frequency graphs or tables continuous probability solutions are expressed via 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 represented in the x-axis and the probability density function represented in the y-axis.

Continuous Probability Distribution

Histogram

Description automatically generated with low confidence

In terms of types of continuous probability distributions, the normal distribution remains the workhorse of the category. In addition to it representing the distribution of many continuous random variables in biology social sciences and business it can also approximate the Poisson distribution as well as the binomial and hypergeometric distribution (Shakil et al., 2010). Additional continuous distributions which will not be covered in this section but are nevertheless important for specific applications and can be referenced in your readings beta distribution, Cauchy distribution, exponential distribution, gamma distribution, logistic distribution, and the Weibull distribution. 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](https://www.calculushowto.com/whittaker-function/) (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 probability distribution and a continuous probability distribution. Consequently, a continuous probability distribution is typically represented using an equation or formula rather than using tables like with a discrete probability distribution (Crooks, 2019). This formula is often referred to as the probability density function. Continuous random variables represent some of the most important measurements of human and business activity on the planet and continuous probability distributions are critical to the understanding and prediction of these variables. However, it is important to remember that given the nature of continuous probability distributions working with discrete distributions is much easier than working with continuous distributions which motivates many researchers to begin their analysis with a discretized version of a continuous variable (Crooks, 2019).

Self-Check Questions

1. Please complete the following sentence:

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.

## 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 but would rather tend to vary and create an average variety that would have outliers, but most breeds would hover around the center of a distribution. This was a principle known as regression toward the mean and was first discussed by Francis Galton who first noticed this in agricultural experiments but then extended the study to heights with parents and their children. P. L. Chebyshev (1821- 1894) was a Russian mathematician, who ascertained that the percentage of observations falling between two distinct values, 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 to the above percentage (Fogarty, 2017). For any population or sample, at least

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

**Normal distribution**

Population parameters define the shape and probabilities of any probability distribution, including the normal distribution. The shape of the normal distribution is defined by two parameters, namely the mean and standard deviation.

Finally, Gauss in 1809 and Adrian independently derived the formulas for the normal distribution. However, Gauss got the credit for this and in fact to this day another name for the normal distribution is the Gaussian distribution (Fogarty, 2017). Since the normal distribution is not a closed-end solution one needs to use integral calculus with much effort to find the true probabilities. Therefore, the hypothetical normal distribution probabilities tables are used to ease the calculations. These probabilities represent the area under the **normal distribution** with 1 equaling the entire area and all if the individual values depending on where the value lies on the distribution in relation to the mean (Fogarty, 2017).

The normal distribution is bell-shaped and is 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 curve which is much skinnier. 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 typical normal distribution.

Typical Normal Distribution

Chart

Description automatically generated

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

Parametric statistics require the data to be normally distributed. These are among the most powerful techniques used in business and the social sciences. If the data are not normally distributed then researchers will have to either transform the data or else, they may 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 deviated from the mean there is a lesser likelihood of 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 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 probabilities and therefore sums to one (Doane, 2016). Interestingly, in a perfectly normal distribution the mean, media 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 which 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. What is the difference between a normal distribution and a standard normal distribution? A normal distribution is determined by two population parameters which are the mean and the variance. These can take on any values but regardless share the same probabilities. Taking this into account a normal distribution can be standardized to have a mean of 0 and a standard deviation of 1. This standardization creates what is referred to as the standard normal distribution (Doane, 2016). The values or raw scores are typically standardized by turning them into z-scores. This standardization procedure enables researchers to ascertain the proportion of value which fall within a specified standard 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 hypothesis 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% will be within two standard deviations of the mean, and around 99.7% will be within three standard deviations of the mean. Therefore, students of statistics often use the 68–95–99.7 rule as shorthand to remember the percentages of observations which fall up to the 3 standard deviations from the mean in a normal distribution (Doane, 2016).

The normal distribution recorded by Gauss is bell-shaped with probabilities representing the area under the curve with 1 equaling the entire area and all if the individual values depending on where the value lies on the distribution in relation to the mean. The normal distribution Is important in the study of statistics because many random variables in the business and personal world are normally distributed. In addition, the central limit theorem states that if one takes a sufficiently large sample with replacement then the distribution of sample means will be approximately normally distributed. This is irrespective of the underlying distribution which at the extreme can even be a uniform distribution. Students' knowledge of the normal distribution can allow them to take advantage of the empirical rule. The empirical rule states that in a normal distribution around 68% of data will be within one standard deviation of the mean, around 95% will be within two standard deviations of the mean, and around 99.7% will be within three standard deviations of the mean.

### Self-Check Questions

1. Which of the following is another name for the normal distribution?

* *Gaussian Distribution*
* Poisson Distribution
* Uniform Distribution
* Dirichlet Distribution

1. Please complete the following sentence:

The normal distribution is bell-shaped and *symmetric* around the mean.

1. Please complete the following sentence:

In a normal distribution the *standard deviation* determines the fatness of the distribution.

1. The total area under the normal curve totals to what?

* 0
* 2
* *1*
* Infinity

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 Francis Galton and developed statistical techniques to validate Darwin’s Theory of Evolution. One of Pearson’s own protégés was William Sealy Gossett who was a gifted statistician hired by the famous Guinness Brewing Company to improve the quality processes at the brewery. Gossett found that the present statistical techniques which 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 to being proprietary to only Guinness Brewery. The communication of this type of information to scholars is primarily done through scholarly journals 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 risk of being continually employed as hypothesis testing which includes the use of t-tests is regarded to be one of the most important discoveries of the 19th century as we will address later in this section and others (Fogarty, 2017). The illustration below depicts a typical t-distribution.

Student’s t Distribution

Chart, histogram

Description automatically generated

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; Tattar, Ramaiah, & Manjunath, 2016; McEvoy, 2018).

**Degrees of freedom**

This concept was first mentioned by Carl Frederick Gauss in the 1880s and brought into modern statistics by William Sealy Gosset in the early 1900s.

### Degrees of Freedom

In applied statistics, the (df) or degrees of freedom represents the number of independent values 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 (typically a positive whole number) are usually represented by taking the sample size and subtracting the number of parameters needed to be calculated (Frost, 2019).

The idea around degrees of freedom is that the the quantiity 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 estimates and 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 when sample sizes which are less than 30 observations this results in researchers having to take into account sample sizes 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 sample sizes exceed 30 observations (Virginia Tech, 1999).

However, the t-distribution is still very similar to the normal distribution. One thing is that it has a precise mathematical definition. Other similar features include the smooth shape of both the normal distribution and the *t-*distribution. In addition, the normal distribution and the t-distribution are both symmetric in that each side of the mean will mirror the other. Both distributions also have a mean of zero. And finally, and also stated above, the ultimate similarity 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).

### Self-Check Questions

1. In what way does the shape of the t-distribution look different from the standard normal distribution?

* Not bell-shaped
* Straight line
* *Heavier tails*
* Right skewed

1. Please complete the following sentence:

The t-distribution is also known as s*tudent’s* t-distribution.

1. Please complete the following sentence:

The t-distribution is used in *t-tests* which are among the most important classes of hypothesis tests.

1. When is the t-test more effective than other tests?

* When sample sizes are large
* For data which has a lot of missing values
* *When sample sizes are small*
* When the data are not normal

## Summary

A t-distribution (aka 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 distributed population. The t-distribution is very similar in properties and shape to the normal distribution. However, the t- distribution is more useful and robust for hypothesis testing when sample sizes are smaller. The standard normal distribution or z-distribution which allows one to perform hypothesis testing using z-scores assumes that students know the population standard deviation. The problem with this assumption is that if students know the population standard deviation, then they probably would also already also know the population mean. The *t-*distribution is based on the sample standard deviation which is a very useful property. The t-distribution is used for one of the most versatile and important hypothesis tests which is known as the Student’s t-test which was originally developed at Guinness Brewery and was brought to the public domain via a collaboration between Karl Pearson and William Sealy Gossett.

# 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.

… explain 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 being traded on the Swiss stock exchange (SIX). Hermann Marx, the CEO, began working with Max Schultz, the CFO to create an annual shareholder report. Hermann and Max had many key performance indicators (KPIs) calculated for the organization 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 contained 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 shareholders. Max had a great suggestion in that any time some of the KPIs displayed a high degree of variance as measured by their standard deviation they would create confidence intervals and report a range of value the population parameter could reside from within. 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% confidence level.” Max agreed with Hermann’s decision.

4.1 Point Estimation

**Point estimators**

A single value or point is used to draw inferences about a population by estimating the value of an unknown parameter in a point estimator.

**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 statistics including means, proportions, medians, and standard deviations (Doane, 2016; McEvoy, 2018).

A point estimate can be useful if a researcher desires a single number to be representative of the entire population. This is only possible where central tendency is present 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 distribution and therefore researchers are only trying 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 consistency and efficiency. Managers measuring performance favor point estimates because they can assign people, process, and businesses a goal based on the measure. However, managers are cautioned that the results may just be the result of random fluctuation without a hypothesis test or using an interval estimate (Doane, 2016; McEvoy, 2018).

There are diverse types of point estimators one can learn more about in the readings. Each of these estimators contains different properties in relation to being reliable estimators of the population. Some of the more popular ones include: bayesian estimators, [minimum-variance mean-unbiased estimator](https://en.wikipedia.org/wiki/Minimum_variance_unbiased_estimator) (MVUE), [best linear unbiased estimator](https://en.wikipedia.org/wiki/Best_linear_unbiased_estimator) (BLUE), m[inimum mean squared error](https://en.wikipedia.org/wiki/Minimum_mean_squared_error) (MMSE), [maximum likelihood estimator](https://en.wikipedia.org/wiki/Maximum_likelihood_estimator) (MLE), median-unbiased estimator, [method of moments](https://en.wikipedia.org/wiki/Method_of_moments_(statistics)) and [generalized method of moments](https://en.wikipedia.org/wiki/Generalized_method_of_moments) ([Lehmann](https://en.wikipedia.org/wiki/Erich_Leo_Lehmann) & Casella, 1998).

A point estimate is used for when an analyst desires to have 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 upon random sampling error. It is always recommended that 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 population or sample. Take for instance a human resource manager who is asked by a candidate about the average salary of the firm. Well, if the CEO and CFO of the firm who are the founder both make $500,000 Euros per year and a random sample of the 20 other employees of the company 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 outliers, or she can still present the mean using an interval estimate covered in the next section.

### Self-Check Questions

1. Which of the following represents a point estimate?

* *A mean of 5*
* A gamma distribution
* A proportion between .5 and .75
* The tossing of a coin 5 times.

1. Please complete the following sentence:

Point estimates are a component of *inferential* statistics.

1. Please complete the following sentence:

A point estimate is utilized if the researcher desires a *single* estimate of the population.

1. What do point estimators extract data from a sample to achieve?

* *The most likely estimate of the population parameter*
* A range of values where the population parameter is most likely to lie
* To obtain the proper sample size
* To ensure that the data are normally distributed

4.2 Interval Estimation

**Interval estimate**

An interval estimator draws inferences from a sample about a population by estimating the value of an unknown population parameter using an interval rather than a single point. With an interval estimator, we attempt to construct an interval that contains within it the actual population parameter with a specified probability.

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 of reporting a point estimate. Examples from the latter include business managers who fear they may be judged on their performance around making a point estimate when a range of estimates would be more appropriate from a statistical perspective than single number to be representative of the entire population. In this case the interval estimate, which is a range of values of which the actual population parameter, has an estimated probability of falling within would be used. 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, interval 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 via reducing the standard error. An example application is supposing we are trying to estimate the income of workers in a factory. Then, 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, Kuhn, & Wojtys, 1998; Altman & Bland, 2011)*.*

The formula for confidence intervals for the mean are depicted below:

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 instead is a categorical variable. Examples of this include political affiliation, gender, or 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, Kuhn, & Wojtys, 1998; Altman & Bland, 2011)*.*

The formula for confidence interval CI for the proportions is depicted below:

where:

= sample proportion

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

n = sample size

Once you calculate a z-score one can look it up in a z-table which is depicted partially below. For access to the full table see (Farber & Larson, 2017).

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

Self-Check Questions

1. Please complete the following sentence:

A *confidence* interval is one of the most popular interval estimation techniques.

Summary

A point estimate is used for when an analyst desires to have 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 upon random sampling error. This variation is then taken into account when calculating an interval estimate as the interval estimate (i.e., confidence interval) will return a range of values of which the population parameter is likely to fall into taking into account random sampling error. It is always recommended that 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 population or sample. An interval estimate can address this all in a single statistic although it's not really a single statistic but instead a range of values but no matter what it is certainly easier to interpret than the former.

## 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 German auto manufacturer the Black Forest Motor Werks known as BFM is anticipating implementing 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 information from the vehicle or interact and adjust certain systems including climate control. The challenge is that the cost of modifying the system across all vehicles is substantial. Hans Brekker, the chief engineer of BFM is discussing with Gerald Meir the chief statistician about how to conduct a test without making a big investment in retooling the plants. The decision was 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 would then monitor the customer satisfaction scores over the next 24 months to see if the new onboard computer interface yielded significantly better results. In order to determine statistical significance, Gerald Meir recommended a t-test be performed on the data. However, he needed to discuss a few things with Hans Brekker including the sample size, the level of significance, the 24-month observation window, and even the logic of using customer satisfaction scores over other options such as net promoter scores. Hand Brekker felt a little overwhelmed by learning about all these new concepts, but at the same timer was happy that this new interface could be successfully tested without making big investments. He reflected up how important hypothesis testing was as a tool available to business organizations like BFM.

## 5.1 Hypothesis Testing when Population Standard Deviation is Known

In hypothesis testing, it makes a big difference whether the population standard deviation is known or unknown. This knowledge determines the probability distribution, which is used to conduct the test. Practically speaking, we almost never know the population standard deviation, but it makes sense to study this since it facilitates a better understanding of the concept of hypothesis testing. Therefore, we will use the next several sections to provide a history, overview, and example of hypothesis testing.

## Origins of Hypothesis Testing

Salsburg (2001) describes a tea party at the University of Cambridge, England where a lady in attendance makes a claim that tea poured into milk can be distinguished from a taste perspective from milk poured into her tea. Most of the members of the party disagreed with the lady on this basis 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 which could satisfy both the lady and her challengers. His test included subsequent randomized presentations of cups of tea to the lady where the milk was introduced first and vice versa with the recording of the lady’s response about the conjecture. This test he proposed was a hypothesis test and the work of Fisher and his contemporaries included Karl Pearson and William Sealy Gossett creates a testing strategy which has been responsible for helping to usher in some of the greatest discoveries known to humankind. So why do we need this type of testing? Well for the simple security that when we make a conclusion about a phenomenon it is not occurring by change and is instead related to the variable in question. In the case of the team the variable is the timing of adding the milk. We needed it in that case because the lady could have been lucky. Fisher’s proposed test, however, was able to measure whether the lady’s chances of getting the number of correct responses was beyond normal chance occurrence. This is the key to hypothesis testing in that associations between two variables may appear to be correlated but in reality, is only the result of randomness. In the following section students will be exposed to the various types of hypothesis tests used in business and the social sciences today. This will allow them to test their own hypothesis about the world around them to see if their conjectures are real or simply due to random variation.

## Generalized Approach to Statistical Hypothesis Testing

Statistical hypothesis testing is a form of statistical inference where 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 methodology for ascertaining the reliability or strength of evidence of insights extracted from the sample in relation to the population from which it was drawn. After all, since it is 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 that the researcher formulates a specific hypothesis based on a hunch 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

With hypothesis testing we are always starting with a research question or research hypothesis which we will then convert into a hypothesis test. This research question should almost always be based on an educated hunch and/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 which is observed from a sample distribution. A step-by-step process is deployed which enables inferences about a population parameter using a sample and taking into account the expected results if the underlying research hypothesis was actually correct (Friedman, 2018).

The five core steps to conducting a hypothesis test are as follows:

1. State the null hypothesis. 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** where one will make a statement claiming a zero difference: a null. The null hypothesis is the one that 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 which is the status quo. It 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 hypothesis is false. The statistical notation for the alternative hypothesis is H1. It is the direct opposite of the null and is what you hope is true as this represents your research hypothesis. Unlike it’s opposite the null hypothesis it never contains the equal sign. H1 ≠ 0 is the statistical language used to represent the null hypothesis (Friedman, 2018). One question which is often asked is why we test for the null rather than the alternate hypothesis which is actually the one we hope to be true. One of the reasons is that in business or the social sciences there are few things that are able to be 100 percent determined 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 outcome and we find enough evidence against this to conclude that are alternate hypothesis 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 which have been conducted over the previous five decades would support evidence 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 deviation. 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 then one would choose the t-distribution which will be covered in the next section. It should be noted that when sample sizes are large (i.e., 30+ observations) then both of these distributions 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 significance level represents the probability of rejecting the null hypothesis when it is true. For example, a significance level of 0.01 indicates a 1% 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 as follows:

Where:

= sample mean

sigma = population standard deviation

n = sample size

The rejection and non-rejection regions are separated by the values of the test statistic. The rejection region represents the set of values for the test statistic that leads to rejection of *H*0. The Non-Rejection Region represents the set of values not in the rejection region that leads to non-rejection of *H*0 (Friedman, 2018).

**The p-value represents 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 value than the observed value of H0 (Friedman, 2018).

As a result H0 will be true.  
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 whole idea of hypothesis testing is to reject the null hypothesis if the sample data does not agree with the null hypothesis. Thus, if observed test statistic is more weighted in the direction of the alternative hypothesis than one is comfortable with from a decision risk perspective than 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 if 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 before is due to the fact that the hypothesis testing process is meant to gather evidence for a research hypothesis versus making a definitive conclsion. Also, students will 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 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 which is whether they are conducting a one-tailed test versus a two-tailed test. A two-tailed test is where we reject the null hypothesis if the t-score for the sample is of a low probability 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](https://statisticsbyjim.com/glossary/mean/) strength of auto parts from an 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 or not the mean exceeds or is lower than the target value (McEvoy, 2018).

Alternatively, there is the one-tailed test, where we would reject the null hypothesis only if the t-score for the sample is a low probability in one direction that we specify before collecting the data. This test makes sense when we have good reason to expect the sample mean will differ from the hypothetical population mean in a particular direction which could be greater than or less than hypothetical population mean. An example of this would be 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 sometimes also referred to as upper and lower tests. When one’s hypothesis has an inequality which 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 compressor life is greater than the original (e.g., 10 years), your hypothesis statements might be

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

When one’s hypothesis has an inequality which 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 3 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 here that the determination of whether it is a left-tailed or right tailed test is determined by the alternate hypothesis(H1) and not the [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/) (Glen, n.d.).

## Hypothesis tests are known to be one of the greatest tools developed in the 20th century. These tests are indirectly responsible for the development of vaccines, life-saving drugs like the cure for Hepatitis-C and even 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 over the population or whether any differences are due to sampling error. All hypothesis tests start with a research hypothesis for what one desires to be tested and this is 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 proves to be very low of the difference happening by chance, then we then have enough evidence to reject the null hypothesis which then supports our alternate hypothesis related to our research hypothesis. The choice of which test to use depends in whether we are able to know the sample standard deviation (t-test) or the population standard deviation (z test).

Self-Check Questions

1. What does the null hypothesis assume?

* *There is no statistically significant difference between the two means.*
* There is statistically significant difference between the two means.
* The degrees of freedom are excessive.
* Random numbers are present in the sample.

1. Please complete the following sentence:

The *two-tailed* hypothesis test assumes the test statistic can move in either direction.

1. Please complete the following sentence:

When the If p-value < α, we would *reject* H0.

1. When the population standard deviation is known, which of the following distributions is used for conducting a hypothesis test?

* t distribution
* *normal distribution*
* gamma distribution
* Poisson distribution

## 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 workhorse of hypothesis testing in business and the social sciences, if we are focused on the differences between two means, is the t-test. In the next few paragraphs, we examine three types of t-tests that are associated with specific research designs: the one-sample t*-*test, the dependent samples t-test, and the independent samples t-test

### 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:

Again, 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 run our calculate our t-score via computer, we will receive as output both the t-score and the p-value. At this point we would implement step 5 above the use our decision logic for hypothesis testing Assuming we decide to test the hypothesis at a 95 percent confidence level than if *p* is less than .05, we reject the null hypothesis and conclude that there is evidence that the population mean differs from the hypothetical mean which is of interest to us. 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 hypothetical mean which is of interest to us (McEvoy, 2018).

If on the other hand we were going to calculate the t-score manually, we could use a table of critical values of t when alpha equal .05 which can be found in any introductory statistical text 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 α is .05. There are also tables for different alpha levels and whether the test is one-tailed or two tailed. Now again applying step 5, the idea is that any t-score any t-score beyond the critical value in *either* direction (for two-tailed tests only) is in the most extreme 5% of t-scores when the null hypothesis is true and has a *p* 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 tet exceeds that value and enterst he region, then the null hypothesis should be refected.

### 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 could be the comparison between a pretest and posttest measuring knowledge of corporate ethics after employees took a new corporate ethics training module. The null hypothesis is that the means calculate 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 in order to make the test more powerful (McEvoy, 2018).

One can frame their interpretation of the dependent-samples t-test as a particular case of the one-sample t-test after the preparation process is completed. This process specifically involves subtracting the two scores for each participant to create a single difference score. With this first step completed, the dependent-samples t-test becomes a one-sample t-test on the difference scores. The hypothetical population mean (µ0) of interest is 0 because this represents 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 now be stated that the mean difference score in the population is 0 (µ0 = 0) and the alternative hypothesis is stated as being that the mean difference score in the population is not 0 (µ0 ≠ 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 vis testing under different conditions in a between-subjects experiment, or they could be preexisting groups in a correlational study (e.g., married and single, male and female). An example application of an independent sample t-test in business would be if a random sample of customers were handled through a live agent or a chatbot and we were to test whether there was a statistically significant difference in their mean satisfaction scores (McEvoy, 2018). The important think to note about these samples is they are chosen independently from one another. The null hypothesis is that the means of the two populations are the same: µ1 = µ2. The alternative hypothesis is that they are not the same: µ1 ≠ µ2. This test should be designated as one-tailed if the researcher has good reason to expect the difference in means to be either higher or lower (McEvoy, 2018).

The *t* 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 follows:

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:

with degrees of freedom df=nA+nB−2.

Please note that the above formula contains the variances which are the squared standard deviations which are contained inside the square root symbol. Also, lowercase nAand nB refer to the sample sizes in the two groups which is 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 specifically confidence intervals. Curran-Everett (2009) discusses the fact that unlike hypothesis tests confidence intervals are a fairly recent phenomenon. Jerzy Neyman first developed them in the 1930s and after this George Snedecor added confidence intervals to his historic *Statistical Methods*. Both the hypothesis test discussed in previous paragraphs and confidence intervals can accomplish the same thing but are often used for different purposes. For example, many researchers will use hypothesis testing when they have a pre-specified hypothesis and significance level and want to do a strict comparison. Alternatively, when they desire to describe the magnitude of an effect (e.g., mean difference, odds ratio, etc.) or when they want to describe a single sample a confidence interval may be more useful (Curran-Everett, 2009).

The choice of which test to use depends in whether we are able to know the sample standard deviation (t-test) or the population standard deviation (z-test). There are also different types of t-tests we discussed which are chosen based on whether there is a single sample or two samples the latter of which we must also determine whether the samples are related or independent. The t-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. An alternative to traditional hypothesis testing is to use confidence intervals which are a component of interval estimations. Many researchers will use hypothesis testing when they have a pre-specified hypothesis and significance level and want to do a strict comparison. Alternatively, when they desire to describe the magnitude of an effect (e.g., mean difference, or odds ratio) or when they want to describe a single sample a confidence interval may be more useful (Curran-Everett, 2009). One final note is that hypothesis testing is not just for means but can also be used to test other statistics including proportions.

Self-Check Questions

1. What is an alternative to hypothesis testing using interval estimation?

* *Confidence intervals*
* Median intervals
* Modal analysis
* Boosting and bagging

1. Please complete the following sentence:

The *dependent* samples independent t-test is used when the two samples are related.

1. Please complete the following sentence:

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).

1. When only the sample standard deviation is known, which of the following distributions is used for conducting a hypothesis test?

* normal distribution
* *t distribution*
* gamma distribution
* Poisson distribution

Summary

Hypothesis tests are indirectly responsible for the development of vaccines, life-saving drugs, and even doubling the life span of individuals in less than a century. Hypothesis tests are known to be one of the greatest tools developed in the twentieth century. The premise if the hypothesis test is simple, in that they are designed to determine whether two treatments 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 for what one desires to be tested and this is 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 proves to be very low of the difference happening by chance, then we then have enough evidence to reject the null hypothesis which then supports our alternate hypothesis related to our research hypothesis. The choice of which test to use depends in whether we are able to know the sample standard deviation (t-test) or the population standard deviation (z-test). This distinction becomes less important if sample sizes are larger. There are also different types of t-tests we discussed which are chosen based on whether there is a single sample or two samples the latter of which we must also determine whether the samples are related or independent. Many researchers will use traditional hypothesis testing when they have a pre-specified hypothesis and significance level and want to do a strict comparison. Alternatively, when they desire to describe the magnitude of an effect or when they want to describe a single sample a confidence interval may be more useful (Curran-Everett, 2009). Hypothesis testing is not just for means but can also be used to test other statistics, including proportions when the data are normal and other statistics like medians with non-parametric tests if 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 UK supermarket chain Anderson’s wants to ascertain whether advertising expenditures 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 answer the critical question of whether Anderson’s is able to attribute their ad spend to increased sales. Helen explained to Alice that it was 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 went on to explain that a popular method of measure in the efficacy of the advertising effort of firms is to build attribution models using simple linear regression models. As a next step Helen agreed to collect data related to the previous 36 months of advertising spend and company sales. The advertising spend would serve as the independent variable and the gross sales would serve as the dependent variable in this study. Hele would plan to run the regression analysis in Excel and then share the results with Alice and her branding and advertising team. Alice was anxiously awaiting the results and hoped that there would be a strong positive relationship between their advertising efforts and sales so she would be able to take it to the executive team to request additional funds to put to work in growing the business.

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

Karl Pearson developed a mathematical rigor around the Pearson product-moment correlation 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 which focused on proving of disproving the theory of evolution Galton study of genetics provided the idea that led to the development of linear regression. In a famous experiment with sweep peas Galton gained the idea for regression by plotting 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 to heredity it also gave Galton the ideas that two variables which were related (the size of mother and daughter sweat peas) could also be predicted. The generalized form of regression analysis was developed later by numerous data scientists (Stanton, 2001). Other contributors to regression analysis include Carl Friedrich Gauss and Adrien-Marie Legendre were scientists who both independently discovered an essential feature of regression analysis which is the method of least squares. Least squares is a statistical optimization technique where the sum of the squared errors are minimized. Both of these scientists used the 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 (Hocking, 1983) and even more recently David Freedman and Andrew Gelman,

## Generalized Approach to Regression

**Linear regression**

This describes a series 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 models (GLM).

Almost all statistical programming languages from R to SAS to even spreadsheet programs including Excel make the calculation of regression analysis see relatively easy while having linear algebra operating as their “engine under the hood” powering the regression equations. While this unit will not be going into the details around the linear algebra its import for students to understand the linear algebraic origins of **linear regression**, so they are better equipped to really take advantage of all the benefits of using regression analysis.

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). For example, this could be trying 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 (Sundaram, 2020)?

Well, remember from algebra our equation for a straight line 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 that 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).

However, in the majority of cases such a perfect line is non-existent. Therefore, we default to searching for another line in this case that when a connecting line is drawn parallel to the y-axis from the data points to the regression line, which measures the error of each data point, the sum of all such errors should be minimized. Sounds really simple and it is by no means incredibly difficult, but we deploy matrix and linear algebra behind the scenes to make this happen (Sundaram, 2020).

Regression along with hypothesis testing paired together are the wheelhouses (key value drivers) of statistical analysis. However, despite all of its popular applications from credit scoring to sales forecasting, regression should not be considered to be a single entity. Rather, regression is a variety of statistical methods all with one thing in common which is the fundamental idea.

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 it may be customer satisfaction score measured on a 0-10 rating scale. The independent variablein a simple linear regression equation which is also known as the predictor variable is what one uses to explain or predict the dependent variable. Continuing along our same these of customer management, this could be for example 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 variable and the x is the Independent Variable. The slope is 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 one comes up with no errors in a model. This should raise a red flag that there may be something wrong with the model. Analysts should try to continually reduce the errors and improve the model (Gray & Gray, 2017).

**Multiple regression**

Solving multiple regression problems by hand requires one tot hink oft he points on an n-dimensional surface in space with n representing the number of independent variables. The shape of this surface depends on the model structure.

This module is attempting to introduce the concept of simple linear regression so that students can 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).

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

Some of the other considerations related to regression analysis is related to the type of dependent variable we are working with. Simple linear regression assumes a continuous 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 customer repays their loan or does not. Since a loan default variable is a binary variable and not a continuous numeric using a simple linear regression that assumes a numeric dependent variable would not be an acceptable statistical. Logistic or Probit Regression in this case would be the correct choices (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 where 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 many times the customer came into the store last year) Poisson and negative binomial regression are two of the best choices for modeling your data (Gray & Gray, 2017).

The examples above are just some of the most popular 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 discussed in the previous paragraphs linear regression is a statistical tool which is used to measure and predict the relationship between two variables one independent (the explanatory variable and one dependent (the predictor variable) 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. Hopefully, the above and previous paragraphs provided a good background to now enable students to learn how to actually develop a simple linear regression model (Yale University, 1997).

Before even attempting to fit a simple linear regression model to the data which has been collected, an analysts 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 to 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 *cause* better employee performance), but this doesn’t take anything 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 quoted to his students that “graphs are your friend to help you understand data”. On that same not 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 probably will not be a productive use of one’s 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 is the [correlation coefficient](http://www.stat.yale.edu/Courses/1997-98/101/correl.htm), 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 in creases so does the y variable. A negative correlation indicates that as the x variable in increases the y variable decreases (Yale University, 1997).

The form of the simple linear regression equation is Y space equals space a space plus space b X , 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* (remember *a* represents the value of *y* when *x* = 0) (Yale University, 1997).

Quality Assessment for Simple Linear Regression

There are several methods to assess the quality of a regression equation. The first is a numeric method and the second represents a graphical method which involved plotting the actual versus predicted values from the regression analysis. The numeric method involves the Pearson’s correlation coefficient which is a statistical tool which ranges from 0 to +1 and -1 and describes a linear relationship between two variables. In the case if 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 linear 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 there is by knowing the independent variable. Of course, the higher 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 tests the regression coefficients for statistical significance. The null hypothesis for a simple linear regression is b = 0 (there is no relationship between variables, the regression coefficient is therefore 0). This can then be tested with a t-test. A 95% 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 sophisticated 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 statistical 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 Y space equals space a space plus space b X , 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 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 measure 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 Pearson’s correlation coefficient squared. The interpretation of this statistic is that it represents the percentage of variance in the dependent variable there is by knowing the independent variable. Of course, the higher the better as long as the model does not overfit.

Self-Check Questions

1. Which component in a regression analysis represents the value when of y when x =0?

* *Intercept*
* Slope
* Regression line
* Error

1. Please complete the following sentence:

Regression analysis when two or more independent variables are involved is known as *multiple regression*.

1. Please complete the following sentence:

Before we begin the process of building a simple linear regression model, it first makes sense to create a *scatterplot* in order to visually assess the extent of the linear relationship between the two variables.

1. Which of the following is one of the key statistics used to measure the quality of a regression equation?

* Z-score
* *R-squared*
* Person product moment correlation coefficient
* F-statistic

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 analysis in Excel and how to interpret the summary output.

Below is the sample data. The research question is: is there a relation between monthly sales (dependent variable) and monthly advertising spend (independent variable). Moreover, can we predict sales if we know the advertising spend?

Sample Advertising Data



### Scatterplot

Before we begin the process of building a simple linear regression model, it makes sense to first create a scatterplot in order 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.

Scatterplot

Chart, scatter chart

Description automatically generated

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

### Simple Linear Regression

We next run a regression analysis on this dataset in Excel using the Analysis Tool Pak add-in to the Data Analysis Tab. Please see the total output below.

Sample Regression Output

**Table

Description automatically generated**

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. R-square or *r²* value in this regression analysis is 0.886 (the square of the correlation coefficient .941), indicating that 88.6 percent of the variation in one variable (monthly sales) may be explained by the other (advertising spend) (Excel Easy, n.d.).

Sample Regression Output – Regression Statistics



### Significance F and P-values

We next need to ascertain whether the results from are reliable and statistically significant. significant), Take a look at significance F (0.000481). If this value is less than 0.05, then you are 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 isn't statistically significant because there isn't sufficient statistical evidence that it's 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.).

Sample Regression Output – Regression Hypothesis Tests

### 

### 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 1151.209. This is valuable information in and of itself. 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 one to ascertain that by using the regression equation above how far away the predicted data points are from the actual data points. For illustration 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.).

Sample Regression Output – Residuals

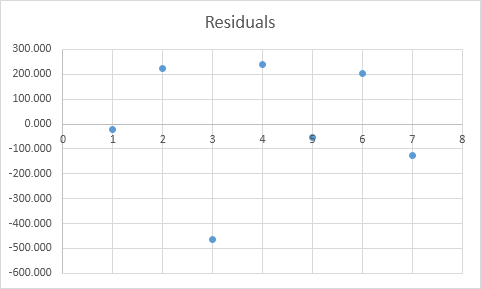


**Residual**

When a residual plot shows a random pattern, it suggests a good fit for a linear model. If the patterns in the residual plot are non-random (i.e., U-shaped and inverted U-shaped), this suggests a better fit for a nonlinear model.

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

Residuals



Outliers and Influential Observations

A point which lies far from a regression line that has been computed for a group of data like the one 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 poor fitting regression. 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.

Monthly Sales by Ad Spend



Please note that points which 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 of the fact that 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 observation tend to be close to the prediction and therefore we do not see any evidence of outliers or influential observations being present (Yale University, 1997).

Lurking or Confounding Variables

Whenever the relationship between two variables is significantly affected by the presence of a third variable, which has not been included in the modeling effort, a lurking variable or confounding variable is present. The presence of lurking variables may manifest themselves in creating non-linear trends which are visible in the relationship between an explanatory 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 and may yield unreliable answers. This practice is known as extrapolation. One example of this would be a linear model which relates employee health risk scores to age for younger employees in their twenties and thirties. 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 one is 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 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 are assuming there are no hidden relationships among the observations and that all of the observations in the dataset were gathered using [statistically valid sampling methods](https://www.scribbr.com/methodology/sampling-methods/).
3. **Normality.** Normality assumes the data are following a pattern that closely approximated the normal distribution.

Linear regression makes one additional assumption.

1. 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 would then make the relationship nonlinear or curvilinear).

Please note that if your data does not meet the assumptions one through three above, you may be able to use a [nonparametric regression such as kernel analysis or regression trees.](https://www.scribbr.com/statistics/statistical-tests/#nonparametric) For violation of assumption four, then nonlinear regression methods including Polynomial regression and Gompertz regression may be potential solutions (Yale University, 1997).

The unique property created by regression analysis lends it to types of useful applications including but not limited to marketing, and operations forecasting, credit scoring, investment portfolio allocation, insurance underwriting and pricing optimization. This module introduced the concept of simple linear regression in a marketing application so that students can conceptualize and easily learn the process. The idea is 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. 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 least-squared algorithm was used to create a line across all points between the two variables, which minimizes the sum of the squared errors for the marketing application we developed. Then the next steps were followed to interpret, assess and apply the equation to the application. Researchers must remember to also always follow several key assumptions when conducting a regression analysis.

Self-Check Questions

1. The attempt to use a regression equation to predict values outside of the original range when the equation was developed is referred to as what?

* *Extrapolation*
* Heterogeneity
* Homogeneity
* Asymptotic

1. Please complete the following sentence:

The most common method for fitting a regression analysis is *least-squares*.

1. Please complete the following sentence:

A *residual* analysis will help us determine how far away the predicted data points are from the actual data points.

1. Which of the following is one of the key assumptions one must make when conducting a simple linear regression?

* Observations in the dataset are less than 350.
* *Linearity between the two variables*
* The dependent variable is discrete.
* There is a curvilinear relationship between the two variables.

Summary

Regression analysis is there to help researchers to ascertain whether one or more variables have a statistical relationship, as well as to predict the future when new observations are made in the independent variable. This tool is therefore 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 space equals space a space plus space b X , 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 quality of our regression analysis. We interpret the r2 as representing the percentage of variance in the dependent variable there is by knowing the independent variable. This unit demonstrated how simple linear regression works in a marketing application using a step-by step process but additional applications of regression analysis including other marketing application include 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 one to ascertain that by using the regression equation above how far away the predicted data points are from the actual data points. Researchers also have to be aware of the presence influential observations and outliers of the data because these both can have an impact on the regression analysis. The presence of lurking variables may also create unwanted non-linear trends when conducting a regression analysis. Finally, there are four key assumptions that researchers must remember to also always follow when conducting a regression analysis. If these assumptions are violated it could lead to biased or unreliable results when interpreting their 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>

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/>

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

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>

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.

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>

Khan Academy. (n.d.). *Statistics and probability*. <https://www.khanacademy.org/math/statistics-probability>

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>

Statistical Solutions. (n.d.). *Common statistical formulas*. <https://www.statisticssolutions.com/dissertation-resources/common-statistical-formulas/>

Statistical Formulas for Statistics. (n.d.). *Some statistical formulas*. <https://1osjhr1kzber15t9781hbik1-wpengine.netdna-ssl.com/wp-content/uploads/2017/04/Statistical-Formulas-for-Statistics-1.doc>

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>

Tutorialspoint. (n.d.). *Statistics—notations*. <https://www.tutorialspoint.com/statistics/statistics_notation.htm>

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).