



## C207 Data Driven Decision Making Notes

Data-Driven Decision Making (Western Governors University)

## Data Driven Decision Making

### Module 1:

1. Data analysis is a process of inspecting, organizing, measuring, and modeling data to learn useful information and make statistically sound decisions.
2. It does mean that managers and leaders must understand the fundamentals of data collection, data analysis, data modeling, and the tools and techniques that are used to inform decisions.
3. The Rise of Analytics
  - a. Driving that decision-making will be quantitative analysis and it will typically focus on statistics.
    - i. Statistics is the science that deals with the interpretation of numerical factors or data through theories of probability. Also the numerical facts or data themselves
  - b. Example: A hospital might want to look at its positive and negative surgery outcomes to compare them to other hospitals or the national average. By going through patient records, they can obtain this data, and then analyze it to compare it with industry benchmarks.
    - i. Benchmarks are standards or points of reference for an industry or sector that can be used for comparison and evaluation.
  - c. Analytics has been defined by Thomas Davenport and Jinho Kim as the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and add value.
  - d. Analytics can help you make decisions based on hard information, rather than guesswork. Analytics can be classified as descriptive, predictive, or prescriptive according to their methods and purpose.
  - e. Descriptive and predictive analytics use past data to project trends in the future.
  - f. Prescriptive analytics, however, can make use of current and future projected data to make suggestions and help direct your decisions.
  - g. Optimization is a prescriptive analytics technique that seeks to maximize a certain variable\* in relation to another.
    - i. Variable is an expression that can be assigned any of a set of values
  - h. Synthesis—the practice of marrying quantitative insights with old-fashioned subjective experience.
4. Big Data
  - a. The use of quantitative analytics is particularly important in Big Data\* decision making. Big Data refers to both structured and unstructured data in such large volumes that it's difficult to process using traditional database and software techniques.
    - i. Example of structured data: Credit card transactions
    - ii. Example of unstructured data: word documents, emails, videos
5. An Example of Analytics – Military Use
  - a. In the military, analytics can be used to measure the quality of the equipment used by soldiers, including vehicles, weapons, or body armor. In product development, analytics can be used to establish parameters that new weapons systems must conform to, or they can assess the efficacy of new training methods.

<b>Descriptive analytics</b>	Depict and then describe the characteristics of what is being studied
<b>Predictive analytics</b>	Use data from the past to predict the future
<b>Prescriptive analytics</b>	Include experimental design and optimization to suggest a course of action

- b. Quantitative analysis of missions can determine how many rations (MREs) will be needed, how much equipment, how many weapons and soldiers. Without quantitative analytics, these decisions would not be informed by hard data, and thus would have a greater chance of failure.
- 6. The Importance of Analytics
  - a. The amount of digital information created and shared in the world has grown dramatically to almost four zettabytes by the end of 2013. A zettabyte is one sextillion bytes. An estimated 90% of the world's existing data was generated in 2012 and 2013. By 2015, many observers expect data creation and sharing to reach almost eight zettabytes.
  - b. Descriptive analytics typically looks at past performance in depicting and describing data and what it means. Predictive analytics lets managers use patterns and relationships in data that to predict business outcomes. They can turn to predictive modeling, forecasting, statistical analysis and other techniques. Prescriptive analytics looks at forecasts and predictions to develop decision options and to recommend a course of action.
- 7. Models of Quantitative Decision Making: Davenport- Kim Three stage Model
  - a. The Davenport-Kim three-stage model\* was developed by Thomas Davenport and Jinho Kim and consists of framing the problem, solving the problem, and communicating results.<sup>1</sup>
    - i. Stage 1: Framing the problem is further broken down into problem recognition and a review of previous findings.
    - ii. Problem Recognition is broken down into:
      - 1. Identifying stakeholders
        - a. It is important that the people to whom you're reporting your results are committed to the project and see the need for the analysis.
      - 2. Focusing on decisions
        - a. Asking what decisions will be made as a result of the analysis is important for three reasons: it helps to identify the reason for the analysis, it helps to identify key stakeholders, and it helps determine whether the analysis is worth doing.
      - 3. Identifying the kind of story you are going to tell
        - a. Although you will be creating your story in stage three (communicating results), you should begin to think about your audience and what kind of story you want to tell with the data.
      - 4. Determining the scope of the problem
        - a. It is important not to get too specific about your experiment at this point, lest you miss an important avenue of investigation.
      - 5. Getting specific about what you're trying to find out
        - a. After reviewing the big picture, focus on a narrow set of data that you will analyze.
    - iii. Stage 2: Solving the problem is the next stage in the model, and is where the mathematical "heavy lifting" takes place. The problem solving stage consists of three steps:
      - 1. The modeling step
        - a. A model is a simplified representation meant to solve a particular problem. For example, a company that is trying to maximize its targeted advertising may create a model. This model could study sales by different age demographics of consumers. Here, sales and age would be the two variables involved in the model; the features being tested are the company's sales and the age of its consumers.
      - 2. The data collection step

- a. This is the part of the project where data is gathered either from primary or secondary sources, and then measured. It is important to recognize the difference between structured and unstructured data. Structured data is data in numeric form that can be easily put into rows and columns. Unstructured data has become more prevalent in recent years and consists of things like text, images, and clickstreams. These things will need to be quantified before analysis can be performed.
    3. The data analysis step
      - a. The goal of data analysis is to find patterns in the data that can then be explained using more sophisticated statistical techniques. The level of analysis will depend on the type of story you want to tell. Remember from the first stage that there are different ways to tell a story. These will be discussed more in the next stage.
  - iv. Stage 3: Communicating Results
    1. Communicating and acting on results is the last stage in the three-stage model. While you may think this is the least important part of the process, it is extremely important if you want your results acted upon.
    2. Some effective visual representations include pie charts, box plots, scatter plots, heat maps and control charts.
    3. When communicating results, or viewing the results of a study, it is important to be wary of the misuse of statistics. Results can be intentionally skewed in order to push a certain agenda.
8. Levels of Measurement – Continuous and Discrete Data
- a. Data is sometimes referred to as either continuous or discrete.
    - i. With continuous data, a data point can lay along any point in a range of data.
      1. Example – age
    - ii. Discrete data can only take on whole values and has clear boundaries. It is not possible to own 3.4 cars; you own either three cars or four.
  - b. Discrete Data Points:
    - i. Nominal - sometimes called categorical data, is used to label subjects in a study. Nominal data is a type of **discrete** data.
      1. Example: Males as 0 and Females as 1
    - ii. Ordinal - is a type of **discrete** data. It places data objects into an order according to some quality. Therefore, the higher a data object on the scale, the more it has of a certain quality.
      1. Example: a third-degree black belt is presumed to have more expertise in karate than a first-degree black belt
  - c. Interval Data Points - is a type of **continuous** data. It has an order to it and all the objects are an equal interval apart, so in interval data the difference between two values is meaningful. You cannot have a natural zero point in interval data, and zero does not represent the absence of the property being measured
    - i. Example: temperature, time
  - d. Ratio data\* is a type of **continuous** data, like interval data. Unlike interval data, ratio data has a unique zero point. With ratio data, numbers can be compared as multiples of one another.
    - i. Example – age
      1. Someone can be twice as old as another person, and it is possible to be zero years old.

- ii. In business, ratio data is common. For example, income, stock price, amount of inventory, and number of repeat customers are all examples of ratio data.

## 9. Reliability and Validity of Data

- a. In statistics (as well as science) measurements need to be both reliable and valid. Reliable data\* is both consistent and repeatable.
  - i. Example: If you were to administer the same test to the same person three times and the scores were similar each time, the test could be categorized as reliable.
- b. Similarly, valid data\* is data resulting from a test that accurately measures what it is intended to measure.
  - i. For instance, if a test reflects an accurate measurement of a student's abilities, it is said to be valid.
- c. Errors: Random Vs Systematic
  - i. All measurements contain some degree of error. This error may be random or systematic.
  - ii. Random errors\* should cancel themselves out over a large number of measurements if they are NOT related to the true score and if there is no correlation\* between the errors.
    - 1. Random errors are errors in measurement caused by unpredictable statistical fluctuations
    - 2. Correlation is the extent or degree of statistical association among two or more variables
  - iii. Systematic errors\* are not due to chance, and although they can be corrected, correcting them takes time and attention to detail.
    - 1. Systematic errors are errors in measurement that are constant within a data set, sometimes caused by faulty equipment or bias.
    - 2. Skewness is a measure of the degree to which a probability distribution "leans" toward one side of the average, where the median and mean are not the same.
- d. Measurement Bias
  - i. Measurement bias can invalidate the results of any study, so it is important not to let bias creep into your experiment.
    - 1. Measurement bias is a prejudice in the data that results when the sample is not representative of the population being tested.
    - 2. A population is an entire pool from which a sample is drawn. Samples are used in statistics because of how difficult it can be to study an entire population.
  - ii. To produce unbiased results the sample tested must be sufficiently random.
- e. Information Bias
  - i. Assuming your sample is properly randomized, the second way bias can enter your model is when data is collected. This is called information bias\* and may occur for a variety of reasons.
    - 1. Information bias is a prejudice in the data that results when either the respondent or the interviewer has an agenda and is not presenting impartial questions or responding with truly honest responses, respectively.

## 10. Concepts of Measurement:

- a. One of the most critical steps in undertaking a research project or quantitative analysis is defining the indicators, or specific measurements, that tell us what a data point represents and what it means for the outcome of the research. In other words, before any measurement can take place, the thing being measured must be defined.
- b. For some variables, we can measure quantitative attributes such as size, frequency, dollar amounts, number of incidents, and so forth. Other variables are more abstract--such as personality type, social

class, or communication style--and their properties need to be converted to something quantifiable, such as a score or category, before they can be measured and analyzed.

- c. Statistical techniques will allow you to manipulate data to better understand the information and make more informed decisions.

## 11. Data Management

- a. Data management\* refers to cleaning and organizing a data set\* that has been collected. Most of the actual data you receive is not ready to be analyzed. Data management is a vital step in the decision making process, and good data-driven decisions rely on clean data.
  - i. Data Management – the management, including cleaning and storage, of collected data.
  - ii. Data Set – a collection of related data records on a storage device
- b. Because humans are involved in collecting and inputting the data, errors will happen. Research suggests, "when humans do simple mechanical tasks, such as typing, they make undetected errors in about 0.5% of all actions. When they do more complex logical activities, such as writing programs, the error rate rises to about 5%
- c. The best way to avoid data mistakes is to spend time checking it, checking it again, and then checking it one last time. The time you'll spend checking will be time you won't have to spend reanalyzing the data.
- d. After collection, your data should be entered into a spreadsheet program or relational database\*. Spreadsheets and databases make it easier to port into and between statistical software packages such as SPSS, SAS, and R.
  - i. A relational database is a database structured to recognize relations among store items of information.
- e. Missing data (an omission error\*) is a very serious data error. Omission errors occurs when something, such as crucial data, is missing. The missing data may be intentional, unintentional, or even a fault of the study.

## 12. Data Quality

- a. Starting with accurate data will give you reliable results when that data has been analyzed. Starting with flawed data will produce questionable analyses.
- b. GIGO (Garbage In, Garbage Out) is the idea that the quality of output is dependent on the quality of input. Originally a computer science term
- c. When looking at data, it is productive to look for outliers\*, observation points (numbers) that are distant from other observations.
  - i. When outliers are detected, we can examine our study techniques and determine whether the number is incorrect and/or can be converted to a correct number. We can also determine whether a figure is an outlier because it describes something that does not belong in the study.
- d. The best fix for faulty data is often to carefully check your work. Having someone else with a set of well-trained eyes examine the results and processes of a study can help identify any problems with the statistics.
- e. Data quality is the state of the accuracy and completeness of data and its suitability to meet the analytical needs of an organization.
- f. Data quality assurance, or DQA, is the process of verifying the reliability and effectiveness of data. Data quality is vital for any analysis, because of the simple, but telling, phrase: "Garbage in, Garbage Out" or GIGO.
- g. Quality data also needs to be useful--data that can shed light on the business challenge involved. Therefore the data must be current and up-to-date if the analysis is going to offer timely insight. The data also needs to be relevant. A database that includes data that doesn't pertain to the challenge or

opportunity being analyzed isn't particularly useful. Finally data becomes valuable for analysis when it is accessible--when, for example, it's in a database that allows for analysis and manipulation.

- h. The Data Quality Cycle focuses on the consistency, accuracy, and relevance of data quality. It begins with the definition of the data needed. It continues with the use of the data where analysis is performed. The next step in the cycle is the validation of the data. Do the results make sense? Are there data errors? Are missing data causing problems? Reviewing the inputs, looking closely at outputs, and retesting the data are some common validation techniques. The final step is making improvements in the data set so that future analysis is more accurate.

### 13. The Uses of Research

- a. Surveys, observations, experiments--these are some of the ways we make sense of patterns and test our assumptions. Research attempts to organize information in ways that answer questions, provide solutions, and make predictions.
- b. We rely on research to help us make decisions and evaluate opportunities for ourselves and our organizations, especially when the problems are complex and involve many different courses of action or interpretations.
- c. While there are many real world applications for well-constructed research, the findings are only as good as the quality of the research design and execution. Far too often, research fails to produce reliable results due to poor research validity and avoidable problem in data collection.

### 14. Research Design

- a. The two main types of research design are observational studies and experimental studies.
- b. Observational studies\* are also known as quasi-experimental studies. An observational study is sometimes used because it is impractical or impossible to control the conditions of the study.
  - i. These studies are conducted in a natural environment where the variables are not completely controlled by the researcher.
    - 1. Example: mystery shopping
  - ii. The best kinds of observational studies are forward-looking, or prospective, and focus on a random group, or cohort.
- c. A prospective cohort study\* observes people going forward in time from the time of their entry into the study.
- d. While observational studies are generally considered weaker in terms of statistical inference, they have one important characteristic: response variables can often be observed within the natural environment, giving the sense that what is being observed hasn't been artificially constrained.
- e. In an experimental study\* all variable measurements and manipulations are under the researcher's control, including the subjects or participants. For example, when studying the impact of price changes on consumers, a researcher can manipulate the price of the product. In such a study, the researcher can control all elements.
  - i. Three elements to an experimental study:
    - 1. experimental units - subjects or objects under observation
    - 2. treatments - the procedures applied to each subject
    - 3. responses - the effects of the experimental treatments
- f. Steps to Setting up a statistical experiment
  - i. Identify the experimental units from which you want to measure something
    - 1. The set of subjects is a sample, or a smaller representation of an entire population\*.
    - 2. Your sample should be chosen at random, in order to ensure that your results describe the population at large. Choosing a sample from a homogeneous group can skew your results.

- ii. Identify the treatments that you want to administer and the controls that you will use, if you will use a control group
    - 1. A treatment is a procedure or manipulation to which you want to expose the subjects to achieve an experimental result. In every experiment there is a treatment group and a control group. A control group is a sample group that is not subjected to the treatment.
  - iii. Generate a testable hypothesis
    - 1. You then generate a testable hypothesis about how the response variable will be affected, run the experiment, and analyze the results.
- g. Validity
  - i. Valid data accurately measures what it is intended to measure. Because valid data is not found by coincidence, those studies that yield valid data can often be repeated many times by different researchers with similar results achieved each time.
  - ii. There are four main types of validity: construct validity, content validity, internal validity, and statistical validity.
    - 1. Construct validity – the validity of inferences that a research study actually measures the construct being investigated
    - 2. Content validity – whether the construct in a research study measures what it claims to
      - a. Can be questioned if the construct is too wide or narrow
    - 3. Internal validity – Occurs when the only variable influencing the results of a study is the one being tested by the researcher
      - a. concerns biases that may find their way into the data that is collected. These may be systematic biases, intentional biases, or self-serving biases
    - 4. Statistical Validity – whether the results of a research study stand up to statistical scrutiny.
- h. Bias
  - i. Assuming you have followed good statistical methodology, you have chosen a completely random sample and you have designed a study that minimizes systematic errors. There are other influences that can introduce bias into your experiment, however.
  - ii. In drug trials, there are generally three populations: the treatment allocator, the participant, and the response gatherer (sometimes the treatment allocator and the response gatherer are the same person). Any or all of these populations can introduce bias into the study if they know who is in the control group and who is in the treatment group. The need to keep participants in the dark about this is self-evident, but subtle attitude changes in the treatment allocator and response gatherer may also have an effect on the study results
  - iii. Blind Study – a study performed where the participants are not told if they are in the treatment group or control group.
  - iv. Double-Blind Study – a study performed where neither the treatment allocator nor the participant knows which group the participant is in
  - v. Triple-Blind Study – a study performed where neither the treatment allocator nor the participant nor the response gatherer knows which group the participant is in.

## 15. Research Standards

- a. There are two major issues involving research standards. The first is the best practices surrounding research standards. The second is ethical: how do we deal with the privacy and the security of the data that we gather and the privacy of research subjects?
- b. The American Statistical Association has established guidelines for statistical practice that address both methodological best practices and ethical issues.



- c. These guidelines provide a roadmap that can guide those conducting research. The ASA guidelines recommend that practitioners use the appropriate statistical methodology and use due caution in drawing causal inferences.

## Module 2: Statistics as a Managerial Tool

### 1. Statistics for Management

- a. Managers can use technology and software such as Excel to handle the math, while still availing them of the useful insight that statistics can offer.
- b. The old truism that "knowledge is power" applies here: statistics can give you a greater understanding of your organization, your customers, and financial markets. Statistical techniques can help you analyze large amounts of "raw data" to extract meaningful insights and estimate probability to predict the likelihood of certain events occurring.

### 2. Employing Spreadsheets

- a. Data-driven decision making today is often performed based on the results of spreadsheet analysis. Modern spreadsheets are quite powerful, allowing data manipulation and the use of functions like linear regression and linear programming.

### 3. Managerial Statistics in Different Sectors

- a. Statistics are used every day in many different areas. They are integral in the decision making processes in business, healthcare, education, military, government, and nonprofit management. The table below provides some of the uses for statistics in these fields as well as examples.
- b. In business, statistics can be used to analyze and make conclusions about large amounts of data, make inferences about a population (that is, customers, vendors, products, or regions), improve predictions, and determine inefficient processes.

### 4. Common Misuses of Statistics

- a. The misuse of statistics can happen due to ignorance or through a deliberate attempt to skew results or to misrepresent data.
- b. Not a truly representative sample
  - i. This misuse occurs when the sample that a statistician chooses isn't truly representative of the entire population he or she will draw a conclusion about. If there are important differences between the sample and the larger population, the conclusion made from the same may not represent the larger population.
- c. Response bias
  - i. This misuse occurs when the respondents to a survey say what they believe the questioner wants to hear. This bias can occur because of the wording of a question.
- d. Conscious bias
  - i. This misuse occurs when the surveyor is actively seeking a certain response to support his or her theory or cause. Bias can occur when the researcher manipulates the phrasing of questions in order to elicit the desired response.
- e. Missing data and refusals
  - i. This misuse occurs when a certain part of the sample gets lost or subjects refuse to contribute to the overall data collection. This can distort the survey data significantly as you could lose demographic segments of the population under study and consequently could arrive at a false conclusion.
- f. Small sample sizes
  - i. This misuse occurs when a sample size is too small to draw inferences from.

- g. Use of wrong tool
  - i. This misuse occurs when a parametric test\* is used while a non-parametric test\* would be more appropriate, and vice versa.
- h. Association and causality
  - i. This misuse occurs when a researcher notices a relationship between two variables and assumes that one variable is the cause of the other. In reality, these variables might both be caused by a separate variable. In this case, they would merely be correlated, which means they show up together. Or there might be no relationship at all.
- i. Training and test data
  - i. This misuse occurs when the same data that's used to form a hypothesis is then used to test that hypothesis. This misuse most often occurs when a small population is being studied and different samples have a lot of crossover.
- j. Unfounded assumptions
  - i. This misuse occurs when an assumption is made that has not been proven.
- k. Faulty operationalization
  - i. Operationalization refers to the development of specific research procedures that allow for observation and measurement of abstract concepts. For example, if a researcher wants to study how new parents feel about their financial security, he/she can operationalize this objective by determining a testable hypothesis; developing a mechanism for collecting observations (for example, a survey); identifying a representative sample (200 randomly selected parents of newborns in suburban towns); asking relevant questions; and interpreting the data received. Each of these research decisions (hypothesis, method, sampling, questions, interpretation) is an example of operationalization.
  - ii. A key aspect of operationalization is defining variables and attributes that adequately represent the concept of the study. For example, to assess attitudes about financial security, researchers might ask respondents to characterize their feelings regarding their financial security as "anxious" "confident" and so forth. These responses can be coded as "1" "2" "3" etc., for analysis. Or they could ask respondents to specify how much money they have in a college savings plan. In both examples, these are measurable dimensions that serve as a proxy for the non-measurable concept the researcher is studying. If the researcher's reasoning behind any aspect of operationalization is faulty, it can result in misleading or irrelevant findings. For example, data showing high amounts in college savings plans may not necessarily correspond to a strong feeling of financial security. The reasoning making this connection is flawed, and leads to inaccurate conclusions.
- l. Lack of blinding
  - i. A lack of blinding can cause bias to occur. Blinding is when researchers place barriers between themselves and subjects in order to ensure that the researchers do not influence subjects' behavior during the experiment. Without blinding, subjectivity can be introduced into the results.

## 5. Evaluating Statistics

- a. When evaluating statistics, there are a number of questions to ask. The following questions should help determine if the statistic could be useful.
  - i. "Was the experiment measuring what the researcher wanted to study?"
    - 1. Researchers should be careful to design experiments so that they measure what is intended to be measured.

2. When designing an experiment, researchers must select variables and attributes that are observable and measurable. At the same time, it is important that they follow a coherent and logical framework to ensure that all aspects of experiment design and method (including research hypothesis, data collection, sample, variables and indicators, data analysis and interpretation) accurately replicate the original research problem.
- ii. Was a causal relationship truly existent?
  1. Determining if an action is causing another action is important in evaluating the conclusions made. A high correlation might only signify they are associated and not that one action causes the other.
  2. Causation – the relationship of cause and effect
  3. The important thing to remember here is that it's not always clear what is causing a certain outcome. A correlation in data is not enough to conclude that a variable affects another variable.
- iii. Was the sample for the experiment representative of the stated population? Were there parts of the sample or population that were missing?
  1. your sample should represent the population by being proportionally distributed through each demographic that might give different responses. If the experiment is applied to a good representative sample, it is still possible that the sample will not respond evenly. This can occur when data is missing or some respondents choose not to respond. Missing data is a serious error and could cause the conclusions drawn from your experiment to be called into question.
  2. While missing data is common in research studies, it is important to understand the source(s) of missing data and whether it is random or non-random. If the missing data is random across the sample, it will not harm the validity of the data except to reduce the sample size. If, however, missing data is non-random (for example, everyone between the ages of 20 and 40 failed to respond), then it will negatively impact the extent to which the sample responses replicate the population.
- iv. Was there a possibility of bias?
  1. An important step in evaluating an experiment is to make sure the experiment is not biased. The goal of an experiment is to collect true data. If bias appears, it will be present in a response or opinion survey or analysis, and the results of your experiment will be compromised.
- v. Was the population used in the experiment the basis for the initial hypothesis and therefore proving nothing?
  1. This part of evaluating a statistic is important when looking at a small population. If you were to form a theory about the 10 biggest companies after knowing a lot about them, you could not conduct an experiment to test that theory with those 10 biggest companies. It is a self-fulfilling prophecy.
- vi. Was the sample size big enough?
  1. It is important when evaluating an experiment to consider the size of a sample. If a sample size is small, people are less likely to believe the theory because the greater the sample size, the greater the precision and, hopefully, accuracy in the results.
- vii. Were any assumptions made that were unproven?
  1. For a theory to have the possibility to be true, all of the ideas on which it is based have to be true.

2. If an assumption is made for the basis of the experiment and the assumption is not proven, the entire experiment is compromised. It is important to follow the logic of how the theory was formed to make sure that no untrue assumptions were made.

## 6. Probability

- a. One aspect of statistical analysis is calculating the probability of events that may happen in the future. When we hear the word probability, we might think of the common probability scenario of heads versus tails when flipping a coin.
- b. When probability equals 0, we know with certainty that an event will not happen. When probability is 1, we know with certainty an event will happen. When probability is .5, or 50/50, we have very low certainty about the future outcome. Low certainty implies greater risk that our decision will be wrong.
- c. Being able to predict where things are going and what might occur informs our decision making and course of action. This might mean taking advantage of possible future opportunities, assessing the likelihood of failure, weighing the risk of doing nothing, or deciding how much to invest in mitigating the effects of unfavorable events or outcomes we see coming.
- d. Analysis of probability is referred to as predictive statistics. A useful predictive tool is trend analysis, which identifies reliable patterns in past data to predict what might happen going forward. If past data trends show strong consistency, we can predict future outcomes with high probability. If the data trend is inconsistent with lots of variance, future predictions will have lower probability.
- e. If a risk analysis conveys a high enough probability of an adverse situation, we can take appropriate risk mitigating actions, based on our tolerance for that particular risk.
- f. Probability also helps us weigh risks and rewards in investment opportunities.

## 7. Introduction to Probability

- a. Probability is the likelihood of an event occurring
- b. If we can calculate what *might* happen in a process or the behavior a group of people (and, in some cases, predict the pattern of events), then it is possible to profit from that knowledge.

## 8. Trials and Events

- a. When determining the probability of different variables, we need previous data. To gather this data, we perform multiple experiments, or trials, and record the results each time.
- b. Events are an outcome that occur and are often represented by a capital letter

## 9. Expressing Probability

- a. Probability is represented by  $P(E)$  which means probability ( $P$ ) of a certain event ( $E$ ) occurring. If  $R$  = Rain and the weatherman says there is a 40 percent chance of rain, then  $P(R) = 0.40$ .

## 10. Calculating Probability

- a. Probability is calculated as the number of ways an event can occur, divided by the total number of possible outcomes.
- b. The opposite of an event happening (i.e. the event not happening) is called the complement\* of the event. The sum of the probability of an event and the probability of its complement is always equal to 1.

## 11. Independent Events

- a. Independent events are those that are **not** affected by other trials or events. For example, if you were to flip a coin once, that first result (either heads or tails) would not have any impact when the coin is flipped a second time — the first event gives no indication of what could result from the second event.

## 12. Complementary Events

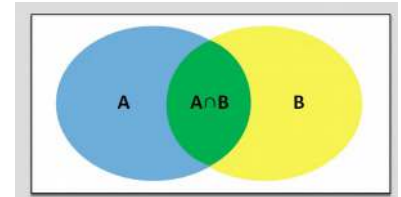
- a. Complementary events are those with two outcomes that are the *only* possible outcomes of that event. For example, flipping a coin and it landing on either heads or tails — those are the only two possible outcomes, so the two events are complementary.
- b. There must also be two defined events in order to have complementary events

### 13. Conditional Probability

- a. Just as events can occur independently, there are also those that occur only in the case of another. Conditional probability is the probability of an event occurring, given that another event has already occurred. These events are considered dependent events\*.
  - i. Dependent events – an event that is affected by previous events

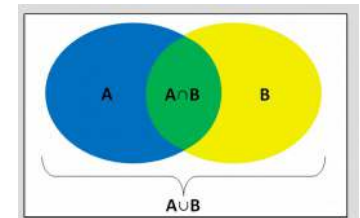
### 14. Probability of an Intersection

- a. The probability of two independent events happening is an intersection\*, and it can be expressed as  $\cap$  in mathematical notation. The intersection of A and B can be written as  $P(A \cap B)$ .
  - i. Venn Diagram – a visual representation of mathematical sets or events
- b. The white area in the Venn diagram represents neither A nor B occurring. For two *independent* events, the probability of an intersection can be calculated as  $P(A \cap B) = P(A) \cdot P(B)$ .



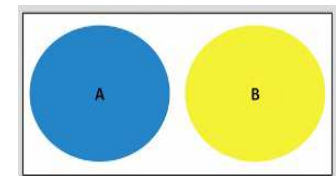
### 15. Probability of a Union

- a. A union\*, written as U, is the chance of, for instance, Bob wearing a black suit OR a black pair of shoes.
- b. A union is the probability of either event happening, including the situation in which both events happen
- c.  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



### 16. Mutually Exclusive Events

- a. If two events cannot both occur, they are called mutually exclusive events\*, or disjoint events
- b. For example, the three statements below are mutually exclusive:
  - i. I was born in January.
  - ii. I was born in March.
  - iii. I was born in May.
- c. Consider mutually exclusive events, A and B. The probability of the intersection of A and B is 0.  $P(A \cap B) = 0$ . Therefore, the probability of union of A and B can be calculated as  $P(A \cup B) = P(A) + P(B)$ .



the

### 17. Probability: Assessing Business Risk

- a. In many cases, such as when there is historical data, it is possible to assign probabilities to estimated values for unknown decision variables, representing the likelihood of that value (or range of values) occurring. These probabilities and the likelihood of positive or negative outcomes can be assembled in a matrix or decision tree to analyze expected payoffs and risk.
- b. Decision Trees - A decision tree is a decision analysis tool that shows a number of options, the paths by which each of these options may be reached, and the possible consequences of choosing each option. A decision tree analysis is designed to establish a logical sequence for decisions, to consider the decision alternatives available, and to evaluate the results they will produce.

### 18. Probability: The Multiplication Principle

- a. In probability, the multiplication principle\* states that to find the total number of outcomes that several events can have, you multiply each individual event's number of possible outcomes by the number of possible outcomes for all of the others.

### 19. Sampling

- a. If you want to test a theory about a large population\* it might be difficult or impossible to test the theory on every individual in the population. This is when sampling\* is used. Sampling is the process of

testing a number of individuals within a population to make a conclusion about the population as a whole.

- b. Making predictions and testing theories about a population from testing a sample is called inferential statistics\*
- c. This differs from descriptive statistics\* which test a population and then make conclusions about only that population.

## 20. Probability: Sampling with and without replacement

- a. Sampling with Replacement: in statistics, a technique used when each piece of the population can be selected more than once
  - i. If sampling with replacement and taking a sample of size  $n$  from a population of  $z$ , there are  $z^n$  possible outcomes
- b. Sampling without Replacement: in statistics, a technique used when each piece of the population can only be selected once.
  - i. If you take a sample size of  $n$  without replacement, there are usually  $n!$  ( $n$  factorial) possible outcomes. The first customer has 6 choices, the 2nd has 5, the 3rd has 4, and so on.
    - 1.  $n! = 6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$

## 21. Probability: Permutations and Combinations

- a. Permutations
  - i. Example: Suppose IC Delight only has one scoop left of each of 8 flavors, and 5 customers come in to buy one scoop each. How many possible outcomes are there now? The first customer has 8 choices, the 2nd has 7, the 3rd has 6, and so on.
  - ii.  $8 \times 7 \times 6 \times 5 \times 4 = 6720$  - Is there an easier way to write this? Yes,  $8! / 3!$ .
  - iii. Permutations are the number of unique ordered possibilities for a certain situation
- b. Combinations
  - i. A combination\* is a sample chosen from a population where the order of the objects chosen does not matter. This means that, in the example at hand, choosing chocolate and vanilla, or vanilla and chocolate, would count as 2 permutations, but only 1 combination. To eliminate these extra permutations and find the combinations of a sample, you need to divide the total number of permutations by the factorial of the sample size.
  - ii. If  $mPn = m! / (m-n)!$ , then  $mCn = m! / ((m-n)!n!)$
  - iii. To solve for the number of permutations, you can make from these flavors we use the permutation formula, expressed as  $4P2$ , .
    - 1.  $4! / (4-2)! = 12$  permutations
  - iv. To solve for the number of combinations, you can make from these flavors we use the combination formula, expressed as  $4C2$ .
    - 1.  $4! / (2! \cdot (4-2)!) = 6$  combinations

## 22. Bayes' Theorem

- a. Bayes' Theorem is a formula that calculates conditional probabilities, important in understanding how new information affects the probabilities of different outcomes
  - i. Conditional probability- the probability of an event occurring given that another event has occurred
  - ii.  $P(A|B) = P(B|A) \cdot P(A) / P(B)$
  - iii.  $P(A|B)$  = Probability that event A happen, knowing that event B happened

The following is the likelihood that any baby is wearing red:

$$P(\text{Red}) = P(\text{Red}|\text{Boy}) \cdot P(\text{Boy}) + P(\text{Red}|\text{Girl}) \cdot P(\text{Girl}) = 0.50 \cdot 0.40 + 0.50 \cdot 0.10 = 0.25$$

From this, using Baye's theorem, we'll determine:

$$P(\text{Boy}|\text{Red}) = P(\text{Red}|\text{Boy}) \cdot P(\text{Boy}) / P(\text{Red})$$

$$P(\text{Boy}|\text{Red}) = 0.40 \cdot 0.50 / 0.25 = 0.80$$

will  
has

## 23. Descriptive Statistics

- a. Descriptive statistics is probably the most common use of statistics. As the name suggests, descriptive statistics uses numbers to describe the relevant characteristics of whatever we're studying.
- b. Measures of central tendency, such as mean, media and mode. Range, or distribution from the smallest number to the largest number. Variability in the data, measured by standard deviation, standard error and sample variance.

#### 24. Khan Academy: Measures of Central Tendency

- a. Frequency: the ratio of the number of occurrences of an event or of certain events compared to the overall possible occurrences of that event or those certain events.
- b. The mode: The mode is the least sophisticated of the three measures of central tendency. Simply, the mode is the single score or value that occurs most often

#### 25. Limitations of Mode:

- a. One advantage of the mode is that it can be used with data from any of the four scales of measurement. However, the mode has a couple of important limitations:
  - i. There can be more than one mode.
- b. The other problem with the mode is that it is not an arithmetic computation. It cannot be used in any of the inferential statistical tools that you will learn later in the course. For that reason, use of the mode as a measure of central tendency is limited.

#### 26. The Median

- a. A somewhat more sophisticated measure of central tendency is the median. The median\* of a distribution is the point at which an equal number of scores fall above and below.
- b. The median is the "half-way" point of the data. An equal number of values in a distribution are greater than the median and less than the median.

#### 27. The Mean

- a. The mean is the most sophisticated of the three measures of central tendency. It is generally known as the arithmetic mean or average. Unlike the median and mode, the mean is influenced by the size of the values in a dataset. Because of this, extreme values (values that are very large or very small relative to the other data points) have a greater influence on the mean than they do on the median or mode. Furthermore, you can only calculate a mean for interval and ratio data.
- b. The symbol for the population mean is  $\mu$ .  $\mu$  is the Greek letter "mu" and pronounced "mew." The symbol for the sample mean is  $\bar{x}$ .  $\bar{x}$  is pronounced "x-bar." The formula for the sample mean is: (see pic)

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

where:

$\bar{x}$  is the sample mean

$\sum x_i$  is the sum of all individual values in a dataset

$n$  is the number of data points in the dataset

#### 28. Properties of the Mean

- a. A deviation score is calculated by subtracting the mean from an individual score. So, the deviation score for the first data point, 44, is  $44 - 50.12 = -6.12$ . The sum of the deviation scores from the mean is always equal to zero. See the table below and notice that performing this calculation for all data points, then adding up the values equals zero.

#### 29. Variance and Standard Deviation

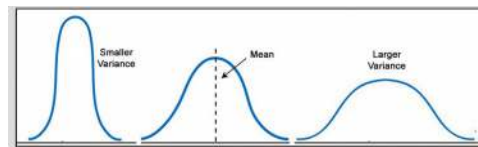
- a. In addition to determining the middle of a distribution by using a measure of central tendency, it is also helpful to know how spread out the data are, or how variable the data are. Measures of variability, the range and the standard deviation, can tell you this.
- b. Variance is a statistical measure of the spread or dispersion of a set of data. Population variance is represented by the symbol  $\sigma^2$  while the sample variance is represented by the symbol  $s^2$ .



- i. The variance\* is a measure of *how spread out data are about the mean*. The closer the data are to the mean, the smaller the variance. If most of the data points are spread out relative to the mean, the variance is larger.

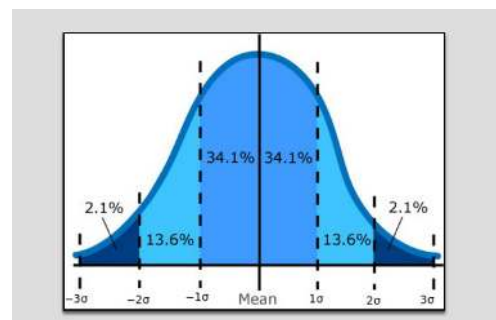
### 30. Khan Academy: Range, Variance, and standard deviation as measures of dispersion

- a. The vertical lines through the bell curves below represent the mean. Notice the difference between the graph of the dataset with the larger variance compared to the data with the smaller variance.



### 31. Calculating Variance and Standard Deviation

- a. Roughly speaking, the standard deviation\* tells you how far, on average, the data points are from the mean. As displayed in the graph below, in a normal distribution the 34.1 percent of the data results will fall between the mean and one standard deviation above the mean. 13.6 percent of the data results will fall between one standard deviation above the mean and two standard deviations above the mean. 2.1 percent of the data results will fall between two standard deviations above the mean and three standard deviations above the mean.
- b. There are two types of variance that each use their own formula: population variance\* and sample variance\*. The standard deviation is always equal to the square root of the variance.



### 32. Calculating Population Variance and Standard Deviation

- a. The variance of a population is denoted by  $\sigma^2$ , and the population standard deviation is  $\sigma$  ( $\sigma$  is the Greek letter sigma). The population variance is formula is as follows:
- c. Remember, population standard deviation ( $\sigma$ ) is the square root of population variance. To calculate a population standard deviation manually, one would take the square root of the population variance formula:

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

where:

$\sigma^2$  = the population variance

$\mu$  = the population mean

$\sum(x - \mu)^2$  = the sum of the squared differences between each data point and the mean. That is, you square each data point's difference from the mean, then add these squared numbers together

$N$  = the size of the population

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

### 33. Calculating Sample Variance and Standard Deviation

- a. The formulas for the variance and standard deviation of a sample (for example, a subset of a population) are slightly different. The variance of a sample is denoted by  $s^2$ , and the sample standard deviation is  $s$ . When calculating the variance and standard deviation for a sample, the  $N$  term in the formula above is simply replaced by  $(n-1)$ :
- b. Remember, sample standard deviation ( $s$ ) is the square root of sample variance. To calculate a sample standard deviation manually, one would take the square root of the sample variance formula:

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

where:

$s^2$  = the sample variance

$\bar{x}$  = the sample mean

$\sum(x - \bar{x})^2$  = the sum of the squared differences between each data point and the mean. That is, you square each data point's difference from the mean, then add these squared numbers together

$n$  = the sample size

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

### 34. The Normal Distribution

- a. Standard deviation is a measure of variance, or how spread out the data is. The standard deviation determines the height and width of the graph.
- b. When the standard deviation is large, the curve is short and wide.



- c. When the standard deviation is small, the curve is tall and narrow.
- d. According to The Empirical Rule, approximately 68.3% of the data points in a dataset will be within 1 standard deviation of the mean. Approximately 95.4% of the data points in a dataset will be within 2 standard deviations of the mean. And almost all (99.7%) of the data points in a dataset will be within 3 standard deviations of the mean.

### 35. Calculating Standard Deviation and Variance in Excel and OpenOffice

- a. The video above introduces how to calculate sample standard deviation and sample variance using =STDEV and =VAR, respectively. Calculating population standard deviation and population variance is also possible with Microsoft Excel and OpenOffice. With all of the other steps remaining unchanged, you can calculate population standard deviation using =STDEVP, and population variance using =VARP as your formulas.

### 36. Graphic Display of Statistics

- a. Graphic display of statistics, also referred to as data visualization, is a necessary tool for communicating and analyzing information. This is especially true in our era of big data, where the data is so voluminous, that we really need to see it graphically to begin to explore it.
- b. The other use, which is gaining in prominence, is as an analysis tool. We use data visualization and charts to explore what the data is telling us as a preliminary step to deeper analysis with other tools. This might involve doing plots of various forms to frame additional questions before we begin to make further sense of it.
- c. Tabulation is a very orderly arrangement of data that allows us to tally information into different categories.
- d. Another useful chart is the histogram. Histograms are tools to show frequency distributions with the height of each bar demonstrating different frequencies in a category. This chart is ideal for making a quick assessment of the distribution of data.
- e. Quadrant analysis is commonly used in competitor analysis, and involves plotting two dimensions on a table.
- f. In the media, we often see a fancy type of data visualization with communication tools called infographics. Infographics are a very powerful way of telling a story that normally would just have numbers.

### 37. Standard Scores (z-scores)

- a. Z-scores will transform different datasets to the same scale. In order to do this, we figure out how far away each individual data point is from its respective mean. In other words, a z-score\* tells us the number of standard deviations a data point is from its mean.
- b. If a data point from the book sale data has the same z-score as a data point from the restaurant spending data, the two data points have the same relative location in their respective datasets.
- c. Because the mean is at the center of the distribution, if the score falls below the mean, the z-score is negative. If the score falls above the mean, the z-score is positive. Once transformed to z-scores, all distributions have a mean equal to zero and a standard deviation of 1.

The formula for z-score is:

$$z = \frac{x - \bar{x}}{s}$$

where:

z = the number of standard deviation units a raw score is from its mean

x = the raw score

$\bar{x}$  = the sample mean

s = the sample standard deviation (also denoted by  $\sigma$  for a population)

Be sure to replace s with  $\sigma$  when dealing with population data.

### 38. Range

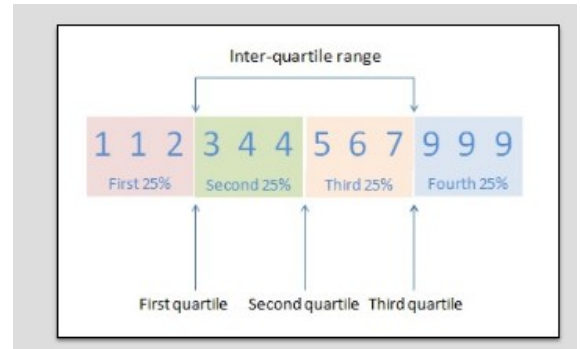
- a. The range\* represents the array of possibilities in which a value can exist, from the minimum value to the maximum value. The size of a range is the difference between the minimum and maximum values.

### 39. Percentiles

- a. A percentile\* is a unit of measurement that gives a value of which a percentage of the population falls below.
- b. It is very important to remember that percentile refers to a percentage of the population and not a percentage of the sum of the values. Percentiles are on the same scale as medians, not averages (means).

### 40. Inter-Quartile Range

- a. A quartile\* is each of four equal groups into which a population can be divided. The inter-quartile range\* measures the difference between the third quartile and the first quartile. To determine the first and third quartiles, order the data from lowest value to highest value. Then separate the data into four equal groups.
- b. The inter-quartile range is an indicator of the distribution of a sample and can also help single out an outlier\*. Outliers are observation points (numbers) that are distant from other observations. It is helpful to identify any outliers and determine whether they should be used.



1. Put the data set in order.  
2 4 6 8 10 12 14 16 18
2. Find the median, or mid-point, of the data set.  
2 4 6 8 **10** 12 14 16 18
3. Identify the median of the lower half of the data set, label as Q1.  
2 4 | **6** 8 **10** 12 14 16 18  
In this case, the median of the lower half of the data set is between 4 and 6, which averages to 5.
4. Identify the median of the upper half of the data set, label as Q3.  
2 4 | **6** 8 **10** 12 14 | **16** 18  
In this case, the median of the upper half of the data set is between 14 and 16, which averages to 15.
5. Subtract Q1 from Q3 to determine the interquartile range.  
 $15 - 5 = 10$

### 41. Box Plots

- a. A box plot can also be known as a box-and-whiskers or hinge plot.

Box plots are used while studying the composition of a data set to examine the distribution.

- b. The top of the line is the maximum value in the set, the upper limit of the box represents the third quartile (75th percentile),
- c. The line in the middle is the median, the lower limit of the box represents the first quartile (25th percentile), and the bottom of the line represents the minimum value of the data set.

### 42. Calculating Quartiles in Excel and Openoffice

- a. To determine the first quartile, type equals sign, quartile, open parenthesis.
- b. Then we select the data by clicking and dragging the cursor to include the whole dataset.
- c. Type a comma at the end of the dataset.
- d. Then select the quartile, in this case type a 1 for the first quartile.
- e. Close the parentheses and click enter.
- f. The third quartile is calculated the same way.
- g. Type equal sign, quartile, choose the dataset, then type a 3 first the third quartile.

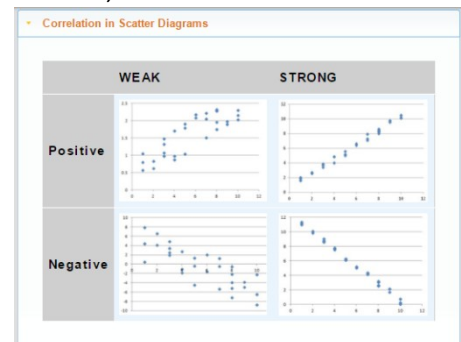
### 43. Histograms and Bar Charts

- a. A histogram\* is a graph that displays continuous data. The vertical bars in a histogram show the counts or numbers in each range. A comparison of the ranges, or a review of the graph as a whole, helps the audience understand the information presented.

- b. While a histogram measures how continuous data is distributed over various ranges, a bar chart measures data that is distributed over groups or categories. For example, a histogram would be appropriate to display how many people fall in various ranges of heights, as height is an example of continuous data. Bar charts, though, would be more appropriate to measure how many people are from each state, as states are an example of discrete categories.
- c. Histograms allow team members and stakeholders to view a significant amount of data at one time, and to see how data is distributed across values and ranges. The histogram's bars represent the values or ranges in the study. The height of each bar shows how many observations or events fall into each range. The shape of the graph illustrates how the data is distributed.
- d. To create a histogram or bar chart, first collect the data using a check sheet\*. For a histogram, decide what ranges, or values, you are going to use and note how the data is distributed. For a bar chart, select the categories for your chart and separate your data into these categories.
- e. A Check Sheet – is a structured form or table that lets practitioners collect and record data in a simple format; by putting marks on a table or image, team members can track and record information about the number, time, and location of events or problems.

#### 44. Bivariate Charts: The Scatter Diagram

- a. One of the most common types of charts is bivariate charts. Bivariate charts have a vertical y-axis and a horizontal x-axis that measure two different variables. A common mistake when drawing bivariate charts is to put the independent variable on the y-axis. This is important because when we look at a graph we naturally look from left to right to see how the dependent vertical variable has reacted over time. Time is the most common independent variable. Because of this, we assume the independent variable is on the x-axis.
- b. A scatter diagram\*, also known as a scatterplot, helps to show potential relationships or correlations between two variables. Data points are plotted as dots along an XY axis, and the concentration or dispersion of these dots shows the strength of the interaction between the variables.
- c. Correlations between variables are very easy to spot on a scatter diagram. If a correlation does exist, the data points on the diagram line up along a curve or straight line across the chart. If the data does not fall along a curve or line, it is likely that no correlation exists between the variables.
- d. Regardless of the correlation suggested, practitioners should be careful not to assume that a correlation among data proves that one variable causes another—it is possible that both of the variables in question are affected by some other factor, or that one variable is a subset of the other.



#### 45. Bivariate Charts: Line Graphs

- a. The other common bivariate chart is a line graph. A line graph\* plots the relationship between two or more variables by using connected data points. Line graphs are very useful where there is time series data to be summarized. They are appropriate where the data are continuous.
- b. Be Wary of Exaggerating Scales!
  - i. By adjusting the X or Y-axis scales you can exaggerate the effect a line chart displays. In the Boeing example, with a Y axis scale (closing price) of \$0 to \$100, we see a gradually increasing price trend over 12 months. However, consider the same data depicted in a line chart where the scale is adjusted to \$50 to \$100 instead of starting at \$0.

#### 46. Proportions

- a. A proportion\* is a type of ratio where the number of observations that are part of a specific group is compared to the total relevant population. It is the frequency of a particular outcome compared to the sum of all of the outcomes (including the outcome that's being focused on).
- b. Proportions can be used when comparing two ratios, frequently concluding that they are equal or unequal. Two ratios are said to be "proportional" if they are equivalent.

#### 47. Probability Distributions

- a. A probability distribution\* is a list of all of the different probabilities of each outcome that can occur. This is often displayed as a graph, table, or formula.
- b. Probability functions, functions that assign probabilities to values of a random variable, determine the information in a probability distribution. A probability function needs to have the sum of the probability values equal to one and also needs each probability value to be greater than zero and less than one.
- c. Uniform probability distribution occurs when any possible outcome has the same probability as any other outcome
- d. There are two types of probability distributions: discrete and continuous.
- e. Discrete data contains distinct values while continuous data can contain any value within a range. The number of working computers is an example of discrete data. One cannot have 2.84 working computers; a computer is either working or is not. The temperature is an example of continuous data. At any moment it could go from 72.01°F to 72.02°F and everywhere between those two temperatures.

#### 48. Cumulative Distributions

- a. A cumulative distribution\* represents the probability that a variable falls within a certain range. Specifically, the cumulative distribution of x measures the probability that a variable is less than or equal to x.
- b. Cumulative distributions, as they are progressing through the data should approach 1.0 or 100% of its data. A cumulative distribution increases as it goes from left to right as it shows the accumulation of information. If the data is discrete, the cumulative relative frequencies are determined by summing each relative frequency to and below the value and then dividing that by the total number data values.

X (outcome)	P (X)	Cumulative Frequency
2 to 9	0.615	0.615
10 to King	0.308	0.923
Ace	0.077	1

of

#### 49. The Central Limit Theorem and Confidence Levels

- a. The Central Limit Theorem\* is the idea that if a great enough number of samples is taken, the means of those samples will be normally distributed around the population mean. As more samples are taken, the sample mean will approach the population mean. Confidence intervals are used to determine the confidence one can have that the true population mean is within a designated range based on information from a sample.

#### 50. The Central Limit Theorem

- a. The Central Limit Theorem states that as more samples are taken, the mean of the set of samples will become increasingly close to the mean of the entire population.

The following is the formula for a confidence level:

(lower interval, upper interval)

$(\bar{x} - (z \cdot s_{\bar{x}}), \bar{x} + (z \cdot s_{\bar{x}}))$

$\bar{x}$  = mean of sample

$s_{\bar{x}}$  (for sample data,  $\sigma_{\bar{x}}$  for theoretical) = standard error of the mean =  $s/\sqrt{n}$  = (standard deviation) / (square root of number of data points)

z = z-score relating to confidence

#### 51. Confidence Intervals

- a. A confidence interval\* is the range around a sample

mean that has a specific probability of containing the true population mean. The "confidence" is the likelihood that a new sample will look like past findings, while the "interval" is the varying range around the existing mean that allows for the different levels of confidence.

- b. The standard error of the mean\* is used to give an estimate of the proximity of the sample mean to the population mean

## 52. Hypothesis Testing

- a. Hypothesis testing is the method of inferential statistics used to make decisions or judgments about population parameters. A common type of hypothesis\* is a statement or claim about a given population.
- b. To test a hypothesis, you must convert the question into a null hypothesis\* and alternative hypothesis\*.
- c. The null hypothesis, or  $H_0$ , is the statement that there is no relationship. For whatever relationship is being tested, the null hypothesis is the statement that the relationship does not exist. For example, if a test is conducted to determine a difference between two means, the null hypothesis will state that there is no difference between these two means.
- d. The alternative hypothesis, or  $H_A$ , is the opposite statement to the null hypothesis. It states that there is a relationship for whatever relationship is being tested. If we conduct a test to determine the difference between two means, the alternative hypothesis states that there is a difference between these two means.
- e. The alternative hypothesis, or  $H_A$ , is the opposite statement to the null hypothesis. It states that there is a relationship for whatever relationship is being tested. If we conduct a test to determine the difference between two means, the alternative hypothesis states that there is a difference between these two means.

## 53. Writing Null and Alternative Hypotheses

- a. The null hypothesis is the statement that is being tested. There are two possibilities after conducting a hypothesis test:
  - i. Reject the null hypothesis.
  - ii. Fail to reject the null hypothesis.
- b. Notice that both of these possibilities pertain to the null hypothesis. The null hypothesis is always the statement that is being tested. The outcome of your experiment is to determine whether the null hypothesis should be rejected. If you reject the null hypothesis, the difference being tested is significant: the difference is most likely not caused by random variation or error. On the other hand, if you fail to reject the null hypothesis, you did not find a significant difference.

## 54. Hypotheses Testing Steps

- a. State the Null Hypothesis and the Alternative Hypothesis
- b. Decide on the Significance Level
  - i. A higher significance level indicates a higher threshold to reject the null hypothesis. To state that there is a significant difference, you have to be more certain that random chance or error is not causing the difference.
  - ii. A lower significance level indicates a lower threshold to reject the null hypothesis. To state that there is a significant difference, you do not have to be as certain that random chance or error is not causing the difference.
  - iii. A commonly used significance level in many research settings is 0.05. A 0.05 significance level means that you state that the results were significant if there is only a 5% chance it was actually caused by random variation or errors. You expect a rejected null hypothesis to be an incorrect decision in only 5% of cases. We will use a 0.05 significance level in most of our examples.

c. Compute the Value of the Test Statistic

- i. One of the most useful kinds of test statistics (that can be used for hypothesis testing in this situation) is the One-Sample t-Test. The One-Sample t-Test can be used to test a null hypothesis concerning a population mean based on statistics from one random sample from the population.

The test statistic for a One-Sample t-test is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where:

$\bar{x}$  = the sample mean

$s$  = the standard deviation of the sample values

$n$  = the number of values in the sample

$s/\sqrt{n} = S_{\bar{x}}$  = standard error of the mean

d. Find the Critical Value and Compare to Test Statistical Value

- i. Ultimately, the goal of the hypothesis test is to make a decision about the null hypothesis. This determination is made by comparing the critical value\* to the test statistic value\*.
- ii. Critical Value is the tipping point between a test statistic value that causes us to reject the null hypothesis and one that indicates we fail to reject the null hypothesis.
- iii. Test Statistic is a value or measure used to summarize what is being tested in a hypothesis test.
- iv. If the absolute value of the test statistic is greater than the critical value, we reject the null hypothesis: there is statistical significance. So, the critical value is the tipping point between where we reject the null hypothesis and where we fail to reject the null hypothesis.
- v. To determine whether the test statistic's absolute value is large enough to reject the null hypothesis, we must find the critical values for the hypothesis test from a distribution table\*. A distribution table corresponds to the test statistic used.
- vi. To find the critical value, you'll need to calculate the degrees of freedom (df) for the test, which is the sample size minus one ( $n-1$ ). In the fitness club,  $df = 30-1 = 29$ .
- vii. We determine your critical t-test statistic's value by finding the table value in the row corresponding to the degrees of freedom (df), and proper column, where  $p$  is the chosen significance level. From the t-statistic table, we locate and identify the critical value.
- viii. There are two possible outcomes:
- ix. **If the test statistic's value is NOT greater than the critical value, we FAIL TO REJECT the null hypothesis.** For a  $t$ -test, this means is that the difference between the sample mean and  $\mu_0$  is not significant. There is no significant difference. There is not enough evidence to conclude that the null hypothesis is false.
- x. On the other hand, **if the test statistic's value IS greater than the critical value, we REJECT the null hypothesis.** For a  $t$ -test, this means is that the difference between the sample mean and  $\mu_0$  is significant. There is a significant difference. Therefore, there is enough evidence to conclude that the null hypothesis is false.

55. Inferential Statistics

- a. Inferential statistics speaks to our need to analyze some characteristic of a very large population without necessarily querying the entire population.
- b. An important aspect of generating valid inferential statistics is ensuring proper sampling of the larger population. Choosing a sample that doesn't contain representatives of all relevant groups within your population can produce misleading results and lead to poor decisions.
- c. To summarize, inferential statistics is an analysis technique that we use to make inferences about the characteristics of a population based on sample data. The sample must be representative of the overall population if inferential statistics will support good data and decisions. An estimation of error measures

how reliable our sample statistics are in generalizing to the population as a whole. To obtain inferential statistics, researchers use a number of statistical tools.

### Module 3: Quantitative Decision Tools

#### 1. Linear Programming

- a. Linear programming\* is a mathematical technique used to find a maximum or minimum of linear equations containing several variables.
- b. Constraints- The state, quality, or sense of being restricted to a given course of action or inaction. An applicable restriction or limitation, either internal or external to a project, which will affect the performance of the project or a process. For example, a schedule constraint is any limitation or restraint placed on the project schedule that affects when a schedule activity can be scheduled and is usually in the form of fixed imposed dates.
- c. Computers and spreadsheet programs allow managers to use linear programming easily and efficiently to help make decisions. Computers today can easily carry out a method known as the simplex method\*, a complicated mathematical method that helps solve linear programming problems.

#### 2. Application: The product mix problem

- a. Managers must often find the optimal mix of products to maximize profit.
- b. A common way to solve the problem in the MindSledge example is to use Solver in Excel or OpenOffice. Solver is a plug-in that automatically pulls all the levers in a linear programming problem to find the optimal solution. The first step is to build the model in a spreadsheet.
- c. As long as a problem has multiple constraints and can be modelled as a series of equalities or inequalities, we can use linear programming to determine a strategy of achieving the most desirable results possible

#### 3. Crossover Analysis

- a. When there are two or more plans or options to consider, crossover analysis\* allows a decision maker to identify the crossover point, which represents the point at which we are indifferent between the plans. With the crossover point identified, it also clarifies which option is better on either side of the crossover point.

#### 4. Break-even Analysis – Example Using the Equation Method

- a. The **margin of safety** is the excess of budgeted or actual sales over the break-even volume of sales (the numbers in green in the chart above.)

Calculating the break-even volume is straightforward using the break even formula:

$$\begin{aligned}\text{Break-even Units} &= \text{Fixed Costs} / \text{Contribution Margin per Unit} \\ &= \text{Fixed Costs} / (\text{Price} - \text{Variable Cost per Unit}) \\ &= \$75,000 / (\$49 - \$24) \\ &= \$75,000 / \$25 \\ &= 3,000 \text{ Units (fans)}\end{aligned}$$

#### 5. Chi-Squared Test

- a. A chi-squared test\* (also written as "**X<sup>2</sup>**" or "**chi-square**") is a common hypothesis test\*. Like the one-sample t-test\*, or the two-sample t test, a chi-squared test is commonly used in statistics to draw inferences about a population, by testing sample data.
  - i. A Chi-Squared test is a hypothesis test that is used to examine the distribution of categorical data.
  - ii. Hypothesis Test is a statistical test used to determine the probability that a hypothesis is true
  - iii. One Sample T- Test is a hypothesis test that is used to compare a sample mean to a known value: often a population mean.
- b. Interestingly, a chi-squared test is employed for categorical data. Categorical data breaks results into categories, like days of the week, or states of the United States of America.



- c. When performing a chi-square test, the **null hypothesis** is the statement that there is no significant difference between the distribution of your data and a specified distribution. Put simply, the null hypothesis states that *the data is distributed as expected*.
- d. All hypothesis tests follow the same four steps:
  - i. **Step 1:** State the Null Hypothesis and the Alternative Hypothesis
  - ii. **Step 2:** Decide on the Significance Level
  - iii. **Step 3:** Compute the Value of the Test Statistic
  - iv. **Step 4:** Find the Critical Value and Compare to Test Statistic Value

#### Chi-Squared Test Formula

The formula for the chi-squared test is:

$$X^2 = \sum (o - e)^2 / e$$

where:

$o$  = the observed value in any given category.

$e$  = the expected value in any given category.

So,  $o - e$  is the difference between observed and expected values in any category.  
 $X^2$  is the sum of  $(o - e)^2 / e$ , for all categories.

## 6. The Normal Distribution

- a. When data tends to occur around a central value with no bias right or left, it gets close to a normal distribution. All normal distributions look like a symmetric, bell-shaped curve.

## 7. ANOVA

- a. Analysis of Variance (ANOVA\*) is a technique used to determine if there is a significant difference among three or more means. Using ANOVA, we see if there is sufficient evidence from sample data of three or more populations to determine whether the population means are all equal, or whether there is a significant difference with at least one of the means.
- b. The null hypothesis\* claims that all population means are equal. For example, if three populations are being tested, the null hypothesis would be  $H_0: \mu_1 = \mu_2 = \mu_3$ .
- c. The alternative hypothesis\* states that not all of the population means are equal. We accept the alternative hypothesis if at least one of the population means is considered significantly different.
- d. An F-value is the test statistic\* that is utilized in ANOVA. As we've seen with other test statistics, our test statistic value and the critical value\* determine whether we "reject the null hypothesis" or "fail to reject the null hypothesis." As always, if our test statistic value exceeds the critical value, we reject the null hypothesis. We would conclude that at least one of the population means is significantly different from the others.
- e. The F-value in the ANOVA output displays the result of the ratio between the mean square of the regression and the mean square of the residual. This gives a relationship between the variability between the groups and within the groups. The greater the difference between the two (making the null hypothesis less likely), the further above 1.0 the F-value will be. This F-value and the degrees of freedom lead to determining if the relationship falls above or below the critical significance level determined by using an F-distribution table.
- f. The "p-value" is the level of significance of a hypothesis test, represented as the probability of a certain event occurring. Because the p-value is less than typical cutoffs for significance (less than .05), it is able to indicate statistically significant relationships. The t-statistic determines whether specific individual variables are significantly related to the dependent variable.

## 8. Different Statistical Techniques

- a. Two of these tools are particularly useful in making predictions and guiding business decisions:
  - i. Regression analysis
  - ii. Time series analysis
- b. Regression analysis is a way to measure how one variable is related to another. Regression analysis identifies a function that describes, as closely as possible, the relationship between these variables so that we can predict what value one variable will assume, if we know the specific value of the other one.



- c. Another forecasting tool is the time series analysis. A time series is a set of evenly spaced numerical observations on a quantitative variable collected over regular time periods. Most businesses keep track of many time series variables, such as daily, weekly, monthly or quarterly figures on sales, costs, profits, inventory, back orders, customer counts, share prices and much more. Forecasts are based only on past values, with no other variables evaluated.
- d. Time series analysis assumes that factors influencing the past and present outcomes will continue to influence the outcomes in the future.

#### 9. Forecasting, Regression analysis, and Quantitative Techniques

- a. Forecasting is one of the most important elements of business decision making. Managers employ three basic forecasting techniques: Judgmental (based on sales, consumer, or management input); Time-Series (based upon data patterns in past data, which includes techniques for random variation, trend, seasonality, etc.) and Associative (based upon predictive or explanatory variables and includes regression.) We're focusing on regression analysis because it is often cited as a powerful technique and managers need to understand it.
- b. *Regression Analysis definition*: statistical method to measure the average amount of change in a dependent variable associated with a unit change in one or more independent variables; considered an associate model as it incorporates the factors (variables) that might influence the quantity being forecasted
- c. Time Series Analysis - forecasting technique that employs a series of past data points to make a forecast
  - i. evaluating patterns in data to make decisions about staffing levels, inventory, etc.
- d. Cluster Analysis - the process of arranging terms or values based on different variables into "natural" groups
  - i. understanding the makeup of an industry's different areas
- e. Decision Analysis - the process of weighing all outcomes of a decision to determine the best course of action
  - i. making decisions, whether personal or professional

#### 10. Benefits and Shortcomings of Regression Analysis and Quantitative Techniques →

#### 11. Regression Analysis

- a. Regression analysis is a statistical technique to measure the link between a dependent variable and one or more independent variables. The goal in regression analysis is to understand the function that explains the impact that the independent variable has on the value of the dependent variable.
- b. Regression analysis is widely used for predicting and forecasting things like sales, labor costs, raw material prices, and so forth
- c. A regression analysis basically draws a "best fit" line through that data and summarizes the distribution pattern. How correlated are the two elements? How much variation is there from a trend? Does the data make sense based on your expectations?
- d. The "standard error" measures the amount of scatter, or variation, in the actual data around the best fit line, or regression function. The standard error also helps us determine a range of predicted values for the dependent variable that we can expect with a high level of confidence.

	Benefits	Shortcomings
<b>Regression analysis</b>	<ul style="list-style-type: none"> <li>- Allows sophisticated analysis of cost behavior and sales forecasts</li> <li>- Provides objective benchmarks for evaluation of reliability of estimates</li> </ul>	<ul style="list-style-type: none"> <li>- Requires 15 or more data points for accuracy</li> <li>- Can be influenced by outliers (unusual data points)</li> <li>- Requires informed analysis</li> </ul>
<b>Time series analysis</b>	<ul style="list-style-type: none"> <li>- Aids decision making by finding patterns in data, such as sales trends</li> <li>- Allows performance and productivity evaluation</li> </ul>	<ul style="list-style-type: none"> <li>- Assumes past data patterns will repeat in future, which may not be true</li> <li>- Key variables may not be captured</li> </ul>
<b>Cluster analysis</b>	<ul style="list-style-type: none"> <li>- Sorts individual data points into different groups</li> <li>- Helps determine target markets</li> <li>- Identifies successful and unsuccessful habits and systems</li> </ul>	<ul style="list-style-type: none"> <li>- Long and expensive process</li> <li>- There are hundreds of potential approaches to take, each specific to a certain situation</li> </ul>
<b>Decision analysis</b>	<ul style="list-style-type: none"> <li>- Determines the decision with the greatest value</li> <li>- Produces a value under certainty, uncertainty, and risk</li> </ul>	<ul style="list-style-type: none"> <li>- Quality of decision is limited to the amount of data available</li> <li>- Does not emphasize the risk of the worst case scenario</li> </ul>

- e. Multiple regression analysis can be applied to these situations. Regardless of the number of independent variables, the goal in multiple regression is the same as the goal in a problem with a single independent variable to find an equation that best explains the relationship between the variables, while minimizing the error.
- f. When carrying out a forecast analysis, we find that there are usually multiple variables that interact with each other. Regression analysis\* is used when multiple variables' quantities relate to each other.

## 12. Dependent vs Independent Variables

- a. The dependent variable\* is the variable whose value depends on the other variables in the equation; typically the cost or activity to be predicted (in the previous example, the dependent variable was home sale price).
- b. The independent variables\* are variables presumed to influence the dependent variable.

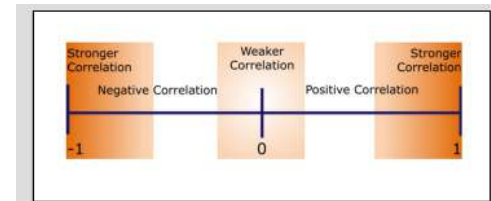
## 13. Linear Regression

- a. A linear relationship between two variables can be measured by its strength. A strong linear relationship indicates that the data will bunch around a straight line, while a weak linear relationship does not.

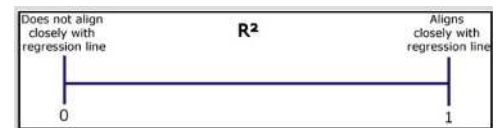
## 14. Correlation

- a. The strength of a linear relationship can be measured with the correlation\* coefficient. A correlation coefficient, a number between -1 and 1, is only useful in measuring linear regression, rather than nonlinear regression.
- b. A correlation coefficient that is close to 0 indicates a weak linear relationship, while a correlation coefficient closer to -1 or 1 represents a strong linear relationship. A correlation coefficient equal to exactly -1 or 1 would be considered perfectly linear.

- c. Negative linear relationships have correlation coefficients less than 0. Positive linear relationships have correlation coefficients greater than 0.

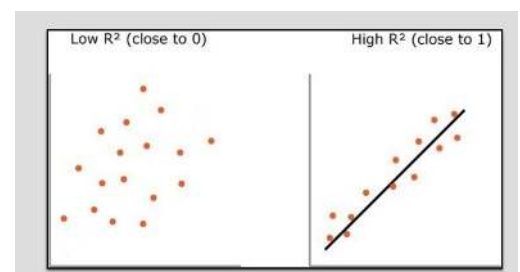


- d. Any regression analysis begins by creating a scatter diagram\* or scatterplot.
- e. We place the independent variable\* on the horizontal x-axis. On the vertical y-axis, we place the variable we are testing, the dependent variable\*. The dependent, response variable responds to the changes of the independent, explanatory variable.
- f. To represent the statistical relationship, we can impose a line through the middle of the points on the scatter plot. This line is called the "line of best fit" or regression line\* for the data and represents a function. The regression line is the function that best minimizes the distance between each data-point and the line, hence the term "line of best fit."



## 15. The R<sup>2</sup> Statistic

- a. In statistics, the term "R-squared" or "R<sup>2</sup>" provides a measure of "goodness of fit." R<sup>2</sup> ranges in value from 0 to 1. An R-squared value close to 1 indicates that the estimation error\* is small and our data closely aligns to the regression line. If there is only one independent variable and one dependent variable, r-squared is the correlation coefficient squared (r = correlation coefficient).



## 16. Standard Error (SE) of Estimate

- a. In multiple linear regression\*, there are ways to measure how well the least squares line fits our data. One important way of measuring the fit of our estimated regression line to our dataset is a process called the standard error of estimate.

- b. Standard error (SE) of estimate\*, denoted  $s_e$ , is the average deviation of the data points from the predictive regression line or curve.
- c. A measure of the accuracy of a prediction driven by a regression model is the standard error of estimate. This differs from the standard error of measurement which measures the amount of variation in the **actual** data around the fitted regression function. Similar to a standard deviation, we know that about 2/3 of the predictions should be within one standard deviation of the actual result. Therefore, a line is more accurate if  $s_e$  is a smaller number.

#### 17. Least Squares Equation

- a. Most spreadsheet programs, such as Excel, can perform a least squares estimation calculation, which determines the best fit line for any of one, or multiple, variables that could affect the dependent variable of our study.
- b. Earlier we discussed how a regression line is the "line of best fit" for a scatterplot. This is determined using an approach called "least squares". In fact, linear regression is often referred to as ordinary least squares (OLS) regression.

#### 18. Calculate the result of a linear regression equation

- a. The regression equation\*, in this case, a line or simple linear regression\*, has the independent x variable affecting the dependent y variable. In other words, the x variables can take on any value, but the y variable is determined by an equation around the x variables.
  - i. So for example:  $y = mx + b$

#### 19. Time Series Analysis

- a. Time series analysis\* is a technique where time is used as an independent variable to assess any influence it may have on an output. Recall that regression analysis allows for one or more independent variables, but requires a single dependent variable. Whether there is one independent variable (simple regression) or more than one independent variable (multiple regression), a time series analysis may be applicable when an independent variable represents time.
- b. Typically, time is measured in sequential intervals of similar duration such as consecutive weeks, months or years. Multivariable regression allows for more than a single input so a test could be conducted, for example, to establish the reliability of (a) defined time intervals, (b) interest rates and (c) the unemployment rate as indicators of a dependent variable such as sales revenue.
- c. As usual we need to discuss a measure of error in these predictions. Complicating this is the concept of confounding variables where it may appear the independent variable studied has a direct impact on the dependent variable being predicted. Sometimes there is a third variable known as the confounding variable which directly affects the independent and dependent variables thus creating an illusion of the relationship between the two variables being analyzed that does not exist.

#### 20. Data Patterns

- a. Trend- A general slope upward or downward over a long period of time
- b. Cyclical- Repetition of up (peaks) or down movements (troughs) that follow or counteract a business cycle that can last several years
- c. Seasonality - Regular pattern of volatility, usually within a single year
- d. Irregularity - One-time deviations from expectations caused by unforeseen circumstances such as war, natural disasters, poor weather, labor strikes, single-occurrence company-specific surprises or macroeconomic shocks
- e. Random Variation - The variability of a process which might be caused by irregular fluctuations due to chance that cannot be anticipated, detected, or eliminated.
- f. Viewing past, or actual, data to construct projections allowing us to make reasonable decisions regarding the future is an ongoing interest to most of us. For example, with a trend upward, a company

might ramp up production and hire more personnel in anticipation of growth, while with a trend downward, it might prepare for contraction and revise its strategy or, even, in extreme cases, exit a business.

## 21. Challenges with Regression Analysis

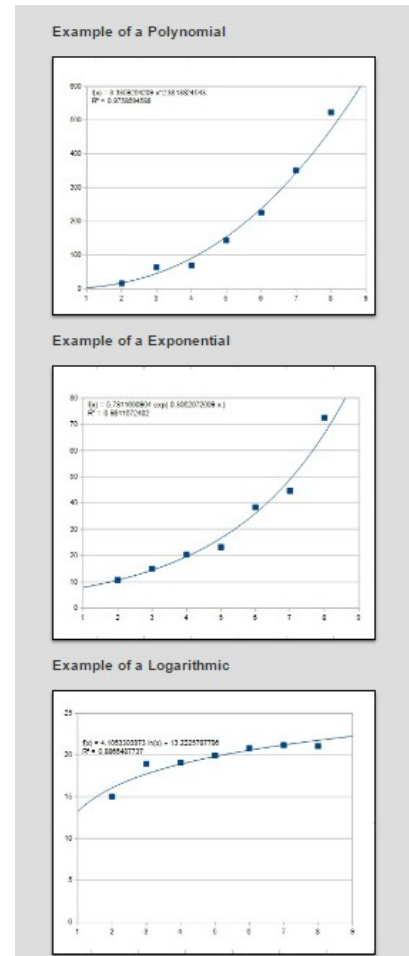
- Previous assignments introduced basic concepts of simple linear regression by considering a case involving a single independent variable\*. Although this is appropriate in some situations, decisions in the real world often involve more than one independent variable. When multiple independent variables are involved, multiple linear regression\* can be used.
- If a dependent variable is affected by more than one independent variable, creating a scatterplot based on only one of those variables might not display a strong relationship. The relationship becomes stronger, and a line will fit the data better, when we account for more of the variables.

## 22. Non Linear Relationships

- Using regression analysis, it may be the case that there is no relationship between or among variables, or perhaps there may be a relationship. Further if a relationship exists it is not necessarily a linear relationship. Thus, our interest in non-linear relationships.

## 23. Outliers

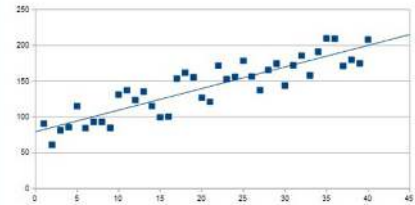
- An outlier\* is an observation point that is significantly distant from the other observations. This could be a valid data point representing a highly atypical measure, or it could be a measurement error. Either way, it is likely to cause the data set to be skewed.
- If you want to ignore these outliers, one possibility is to determine if the outlier is greater than Cook's Distance\* and therefore determine if it's possible to disregard it. Because outliers can represent valid and severe potential outcomes, it is important to determine the source of an outlier.
- Cook's Distance – an estimated distance that is determined by the influence of a single data point on the results of regression analysis
- Multiple linear regression\* is an analysis of how multiple independent variables affect one dependent variable\*. Regardless of the number of independent variables, the goal in multiple regression analysis is the same as in simple linear regression: to find a best fit line within our data and use it to make predictions.
- Logistic regression\* can be applied when the dependent variable is a categorical, binary variable, such as male/female, dead/living, gas/electric, etc. A set of data can follow a trend without that trend being linear. The data points can follow any nonlinear trend, such as a curve or series of curves. Like linear regression, logistic regression can be based on one or more independent variable
- Autocorrelation\* occurs when a given data point on a time series analysis\* is affected by a previous data point for that time series.
- In an ordinary regression analysis we assume that errors are independent from one another. To determine if autocorrelation is present in a regression analysis, we can determine the Durbin-Watson



statistic\* of the dataset. It is important to remember to check for autocorrelation when the dataset under analysis is a time series.

- h. Autoregressive Error Correction produces a superior regression analysis compared to ordinary regression analysis because it takes autocorrelation into account.
- i. Homoscedasticity\* occurs when all of the random variables have the same general finite variance. In other words, the data points in a scatterplot stay approximately the same distance away from the regression line throughout the entire dataset. For ordinary least squares (OLS) regression analysis, homoscedasticity is assumed.
- j. When heteroscedasticity\* occurs, the random variables have an unequal spread of variances. In other words, the data points in a scatterplot tend to be spread to varying distances from the regression line depending on the location of the data point on the line.

Homoscedasticity Example



## 24. Cluster Analysis

- a. Cluster analysis\*, also known as segmentation, is the process of arranging terms or values based on different variables into "natural" groups. Most often with cluster analysis, these terms or values are survey responses from people.
- b. Cluster Analysis has a number of steps:
  - i. Determine the problem by selecting the variables you want to study
  - ii. Select a distance measurement between the values (there are a number of options here).
    - 1. The possible methods for this are the Squared Euclidian distance, the Manhattan distance, the Chebyshev distance, and the Mahalabonis distance.
  - iii. Decide on which clustering procedure to use (this is where a large variation in approaches to cluster analysis occurs).
    - 1. There are two main categories of cluster procedures: K-means clustering and Hierarchical clustering. K-means clustering is used when you are sorting the clusters into groups of values where the values are closest to their clusters average location. The other main category of cluster procedures is Hierarchical clustering. This is used when there is a building of a hierarchy of clustering.
    - 2. Distribution-based and density-based clustering are two other categories of clustering that can be used.
  - iv. Decide on the number of clusters
    - 1. If there is no pre-determined number (k) of clusters, a rule of thumb is to take the square root of half the number of data points ( $k = \sqrt{n/2}$ ). A data point is a single piece of data, usually representing an individual that has a certain answer to a question (people that drink at least x amount of soda, insects that fly, people that watch a certain tv show, trees that grow to at least x feet, etc.). The variables (possibly in the x, y, and even z axes) would be dependent on the individual's responses to other questions (age, ethnicity, weight, level of education, origin location, type of tree, etc.). There are also a number of other ways to determine the number of clusters.
  - v. Map the values into each cluster
  - vi. Make conclusions about the clusters
  - vii. Determine the validity and reliability of the analysis

## 25. Decision Analysis

- a. Decision Analysis\* is the process of weighing all outcomes of a decision to determine the best course of action. This is for any situation a person faces, whether personal or professional.
- b. Under certainty, each action has known outcomes. The outcomes are weighed and the best action should be chosen.
- c. Under uncertainty, there are unknown probabilities of the outcomes of different actions. This is a more common situation than a decision under certainty. Because the probabilities are unknown and there are multiple outcomes, the decision has to be based solely on the possible gains or losses from different actions.
  - i. The "minimax" is determined when the opportunity loss is calculated. The minimax procedure says to choose the option with the least opportunity loss.
  - ii. The "maximin" is determined when the greatest minimum is calculated. The maximin procedure says to choose the option with the greatest minimum.
  - iii. The "maximax" is determined when the maximum payoff is determined. The maximax procedure says to choose the option with the greatest payoff.
- d. Under risk, there are known or estimated probabilities of the outcomes of different actions. Knowing the probabilities of the different outcomes allows for the calculation of an actions' worth. To calculate the expected payoff for an action, you multiply the outcomes' payoff by the probability of that outcome; then all of these together for each outcome.
- e. A more-complex quantitative analysis approach involves a tool called a decision tree\*.
  - i. A decision tree is a decision analysis tool that shows a number of options, the paths by which each of these options may be reached, and the possible consequences of choosing each option. A decision tree analysis is designed to establish a logical sequence for decisions, to consider the decision alternatives available, and to evaluate the results they will produce.
  - ii. The probabilities for all the options should total 100%.
- f. With uncertainty and risk, it might not be a good decision to follow this decision analysis and go with the maximum profit or maximum expected payoff. This is important because it is possible that, although those options might theoretically give the highest average outcome, the losses that could be involved might be too great that the action might not be worth taking.

## 26. Simulations

- a. A simulation\* attempts to emulate a real process or system through an imitative model. This allows considering problems that may not lend themselves to direct experimentation and helps managers make decisions. Common simulation tools include what-if analysis\*, and Monte Carlo simulation\*.
  - i. A what-if analysis is a form of simulation analysis that involves selecting different values for the probabilistic inputs in a model and then computing the possible outputs.
  - ii. A Monte Carlo simulation is a process which generates hundreds of thousands of probable performance outcomes based on probability distributions for cost and schedule on individual tasks. The outcomes are then used to generate a probability distribution for the project as a whole.
  - iii. Simulation models\* use computer-based programs to predict the behavior of a system. The simulation acts as a similar but simpler model to represent a situation or problem in order to analyze possible outputs. It will include observed data as well as randomly generated, predicted data.
- b. Mathematical expressions are used to describe the relationship between inputs and outputs in a simulation model. There are two types of inputs:
  - i. Controllable inputs\*: Inputs that are directly controlled by the company or individual. An example would be the decision by a company to invest in building a new production facility

- ii. Probabilistic inputs\*: Inputs that are outside the direct control of the company or individual and take on different values. An example would be the costs of raw materials used in the new production plant the company had decided to build
- c. What-if analysis is a form of simulation analysis that involves selecting different values for the probabilistic inputs in a model and then computing the possible outputs.
- d. A Monte Carlo simulation is a problem-solving technique used to approximate the probability of certain outcomes by running multiple trial runs, or simulations, using random variables. It lets us model situations that present uncertainty and run them thousands of times on a computer.
- e. While there are several different ways to run a Monte Carlo simulation, there are five basic steps:
  - i. Decide the probability distribution of important variables.
  - ii. Calculate the cumulative probability distribution for each variable.
  - iii. Decide an interval of random numbers for each variable.
  - iv. Generate random numbers.
  - v. Run a series of trials and determine simulated value of the actual random variables.
- f. Simulations have advantages and disadvantages, as summarized in the following chart.
- g. Advantages of Simulations include:
  - i. allows analysis of large, complex, uncertain systems
  - ii. permits studying systems behavior without disruption of actual operations
  - iii. can compress a time frame and enable assessment of effects of a change that take place over several years
  - iv. can employ a large number of probabilistic inputs
  - v. allows the consideration of "what if" scenarios
  - vi. can investigate scenarios that might prove dangerous in real life (such as modeling for health care or manufacturing)
- h. Disadvantages of Simulations include:
  - i. can be expensive and time-consuming to develop and execute
  - ii. dependent on realistic, accurate input
  - iii. can require user to have deep understanding to generate conditions and constraints
  - iv. can be difficult to interpret simulation results
  - v. dependent on assumptions (such as probability estimates) that may turn out to be incorrect

## 27. Break-even Analysis

- a. Break-even analysis is a specific type of crossover analysis. We use break-even analysis to analyze one process by comparing revenue to cost. The break-even point, which is our crossover point between cost and revenue, is where revenue equals cost, profit is zero, and we "break even."

## Module 4: Quality Management Basics (Statistical Process Control)

- 1. Quality management and statistical process control
  - a. Quality management helps ensure that a company's products and services deliver value to customers and stakeholders.
  - b. Statistical software packages, automated data collection systems, and real-time management dashboards are available to support these data collection and quality management initiatives.
- 2. Quality Management Principles
  - a. Quality Management is a component of a business management paradigm that focuses on product/service quality and the means to achieve it.
  - b. The International Organization for Standardization (ISO) has developed several quality management principles to guide practices and to help you frame your quality management activities. By using these



principles as you implement your quality processes, you'll ensure that you take a holistic view of quality management and maximize its effectiveness.

- c. A focus on customers - Organizations should work to understand customers' current and future needs and strive to meet or exceed those needs. This focus on customer satisfaction will lead to increased loyalty from customers, which in turn will lead to an increase in revenue and market share as customers return for repeat business.
  - i. Lean processes focus on eliminating anything that does not add value to customers or satisfy the customers' needs.
  - ii. Plan-Do-Check-Act Cycle- This cycle can be used to work with the customer to create the best possible product available.
- d. A commitment to strong leadership - Quality management activities and programs require strong leaders who will ensure that teams are disciplined in their approach and remain aligned on clear goals and objectives. Having a strong leader to guide activities will increase the team's understanding of its objectives and minimize miscommunication among project participants.
  - i. Tree Diagrams and Process Decision Program Charts are two tools that are useful in strong leadership.
    - 1. Process Decision Program Charts are a type of tree diagram that breaks down a project into different activities and helps mitigate risk.
- e. Engaged Colleagues - People at all levels of the organization must be fully engaged and involved in the pursuit of quality to maximize its benefits to the organization.
  - i. The Interrelationship Digraph is useful when engaging colleagues because it shows the far-reaching impact that an individual or a team can have on the other different areas within the company's operation.
- f. A focus on process - When resources are looked at and managed as a process, results are easier to manage and achieve. A process approach also helps to create consistent, predictable outcomes and uncovers opportunities for improvement.
  - i. A **Supplier-Input-Process-Output-Customer (SIPOC)** diagram is very useful when focusing on the process of managing resources. An SIPOC diagram will show all of the elements that can influence a process before the process has started.
- g. A systems approach - When adjustments are made, their impact on the whole system should be analyzed because a change in one part of the system may affect other parts as well. This systems approach will help to ensure consistency and efficiency among organization-wide activities.
  - i. **Interrelationship Digraph**- useful in understanding the possible impact of a change
- h. A commitment to continuous improvement - Every individual in the organization should be committed to increasing their skill and performance on a regular basis
  - i. The **Plan-Do-Check-Act Cycle** is an important tool when committing to continuous improvement. This cycle can be used to create the best possible process or product. **ISO Certification** can also be useful for a company as it states that a company is dedicated to quality practices and is continually working to ensure that it is producing at the highest level possible.
- i. Dedication to fact-based decision making - By ensuring that decisions are based in fact, organizations can be sure that options are chosen because they are best for the project and organization.
  - i. A **prioritization matrix** helps a team prioritize their options statistically before determining the best possible fact-based outcome.
  - ii. **Six Sigma** can also be useful making fact-based decisions when looking at quality because it measures the performance and determines areas that need improvement.
- j. A collaborative relationship with suppliers



- i. A relationship with suppliers that benefits both parties will encourage interaction and a sharing of expertise that may not otherwise be present.

1. The **Network Diagram** can be a useful tool for supplier relationships because it defines what is needed at the beginning of the process.

### 3. The Plan-Do-Check-Act Cycle

- a. The Plan-Do-Check-Act cycle\* (or PDCA cycle as it is commonly known) is a four-step method for testing hypotheses and solving problems. The cycle is based on the scientific method—develop a hypothesis, run an experiment, and analyze the results—to help gather information and test options before implementing changes on a large scale.
  - i. the Plan step, where you identify a problem and develop plans to solve the problem
  - ii. the Do step, where you run an experiment to see if your plans will work on a small scale before you implement them on a larger scale
  - iii. the Check step, where you analyze the results of your experiment and decide if they can be improved in any way
  - iv. the Act step, where you enact the change on a larger scale, making it a part of normal operations

### 4. Quality Control vs Quality Assurance

- a. Quality management\* involves both quality control\* and quality assurance\* but these terms can be confusing because they are often used interchangeably.

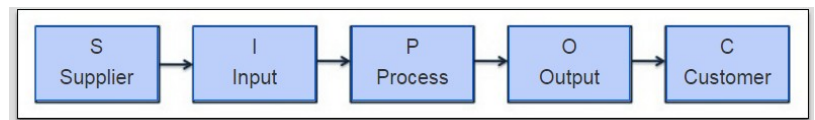
There are, however, distinct differences between them, as shown in the following table:

	Quality Control	Quality Assurance
<b>Focus</b>	Uncover defects so they can be fixed	Prevent defects from occurring
<b>Purpose</b>	Assess performance and recommend corrective action	Assess capability and recommend preventive action
<b>Level</b>	Basic—recognize problems so they can be fixed	Advanced—understand the intricacies of the system and predict outcomes
<b>Major Activities</b>	Inspection and repair	Training
<b>Change Response</b>	Reactive—take action once the problem has occurred	Proactive—take action before the problem can occur

- b. Quality management is not an "either-or" proposition—you don't have to choose between quality assurance and quality control. Instead, you should strive to incorporate both into your quality improvement processes.
- c. Quality control uses observed data to determine the changes that have to be made to a process or procedure. Quality assurance uses projected data to make the best possible model or procedure before that model or procedure is applied.

### 5. SIPOC (Supplier-Input-Process-Output-Customer)

- a. One way to ensure that you are viewing your process as a whole is to create a SIPOC diagram\*. A



SIPOC diagram allows you to see

all of the elements that could influence the process before work is begun. It helps you define the boundaries of your operations by providing a high-level view of the complete process, from supplier to customer.

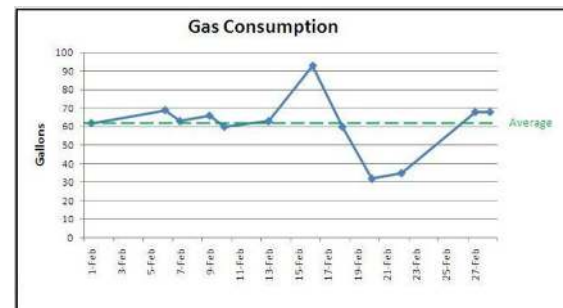
- b. *Suppliers* are the people or organizations who will provide the services or materials that the process will use to create results. Suppliers can be listed as generic roles or as specific individuals if that information is available.
- c. *Inputs* are the services or materials that suppliers provide. Inputs can be recorded as raw materials, service activities, or anything that will trigger processes to start.

- d. *Processes* are high-level descriptions of the work that will produce the results. Processes are often listed in a "verb + noun" format (such as "Create text") and may be shown as a flowchart to illustrate the order of the process steps.
  - e. *Outputs* are the results that will be produced. Outputs may be described as tangible products or as services provided.
  - f. *Customers* are the intended users of the outputs that will be produced.
  - g. When creating a SIPOC diagram, be sure to involve team members and key project personnel in its development. Asking key project participants to help you create the diagram will enhance your ability to view the process from multiple angles and ensures that important information is not missed.
6. Statistics, Metrics, and Quality
- a. Quality management practices often rely on measurements and statistics to quantify information and evaluate results. These measurements and standards help project participants look at processes objectively and provide a simple way to gauge improvements and progress.
  - b. Some quality management methodologies use statistical process control\* (SPC) to help teams monitor work and ensure that processes work to the best of their ability. SPC relies on metrics\* to illustrate results and to analyze the root cause of any deviations from plans.
  - c. By employing SPC methods, a team can reduce the need for inspection by ensuring that quality is "built into" the process instead of "added on" to the end of the process. By making quality improvement activities an integral part of a process, SPC prevents mistakes from being incorporated into an entire batch of a product and also reduces the rework that would be needed to fix the product before it could be released to the customer.
7. Sampling
- a. Metrics help teams measure process results but it may be impractical (or even impossible) to measure every result that a process produces. Testing every product may be too expensive or time-consuming or may destroy the inventory you are producing.
  - b. To resolve this problem, teams often employ sampling\* techniques to help them measure results. Sampling involves choosing one (or several) of the outputs generated from a process as representatives of the entire group. Tests are then run on these samples and findings are extrapolated to represent the entire group.
8. Attribute and Variable Data
- a. Quick data collection or decreased cost may allow you to collect attribute data\*, while a demand for more-specific data would require that you collect variable data\*.
  - b. Attribute Data shows whether a result meets a requirement or not (yes/no)(pass/fail)
  - c. Variable Data shows how well a result meets a requirement, often shown on a scale or as a rating
  - d. Variable data provides more information than attribute data but often takes more time and effort to evaluate and analyze.
9. Common Cause Variation vs. Special Cause Variation
- a. Variations occur in all processes but it is important to distinguish between the variation that occurs naturally as part of the process and an abnormal variation that affects results.
  - b. These common cause variations\* are accepted as part of the normal process because they fall within the amounts that users will tolerate.
    - i. Common cause variations are variations that occur as a natural part of a process
  - c. Special cause variation is abnormal variation that is not a natural part of a process
    - i. To eliminate this special cause variance, the organization would have to "fix" the system by eliminating the cause of the abnormal results in the process or system.
10. Control Limits

- a. There are expectations for the output\* of any system or process. The product is expected to have certain characteristics.
- b. These constraints of a process's output are called control limits\*. The upper constraint of a process is known as the upper control limit\*, while the lower constraint is the lower control limit\*.
- c. An upper control limit and a lower control limit are usually equidistant from the mean. Because they are the same distance from the mean, these control limits are usually designated as a certain number of standard deviations from the mean. The most commonly used distance is  $3\sigma$ : the upper control limit being three standard deviations above the mean, while the lower control limit is three standard deviations below the mean.

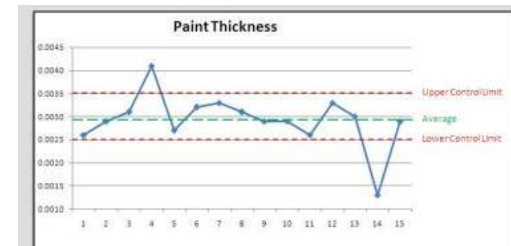
#### 11. The Run Chart

- a. A run chart\* is a simple way to illustrate performance measurements over a period of time. Team members collect measurements and plot them as data points on a graph, then connect the points to show trends or aberrations in performance. These trends or aberrations can be investigated to see if corrective action needs to be taken to address root causes or problems.



#### 12. The Control Chart

- a. A control chart\* is a modified run chart—it shows the performance of a process over time but it also includes limits or constraints that a process should not exceed. These control limits\* track the maximum value (the upper control limit\*) and the minimum value (the lower control limit\*) that the process must perform within.
- b. Carefully monitor any points that approach or exceed the warning limits but have not yet exceeded the control limits, as they may be indicators of upcoming problems. Be especially careful to address any points that fall outside of the control limits.
- c. Control charts are especially helpful in distinguishing special cause variation\* from common cause variation\* due to their increased focus and more-critical evaluation of data. By providing more-refined data, control charts will alert you to potential problems and "hidden" issues that may not otherwise be evident.

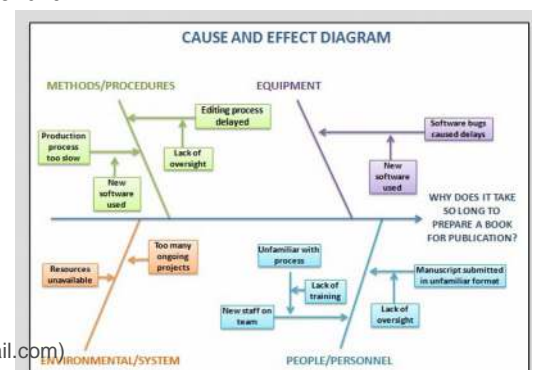


#### 13. Rules for Interpreting these Charts

- a. If your run chart or control chart shows any of the trends or aberrations listed below, be sure to analyze your process for special cause variation:
  - i. 6 or more increasing data points in a row
  - ii. 6 or more decreasing data points in a row
  - iii. 8 or more points in a row that alternate up and down
  - iv. 9 or more points in a row on one side of the center (average) line
  - v. Any point outside of the control limits on a control chart

#### 14. The Cause – and –effect Diagram

- a. A cause-and-effect diagram\* helps project participants systematically uncover sources of problems so they can be examined for similarities and correlations that point to the fundamental reason(s) for the problem. It creates a







hierarchy of the primary and underlying factors that cause an event or problem, allowing teams to drill down to find the problem's root cause.

- The cause-and-effect diagram is often called a fishbone diagram\* because, as team members brainstorm ideas and continue to dig deeper into a problem, they record these ideas as smaller and smaller branches off of a central connecting line.
- There are many category headings that you can use but the most common ones are based on 1) the methods or procedures used, 2) the equipment (tools and machinery) utilized, 3) the system or environment the project or process is happening in, and 4) the people involved in the work.

## 15. The Flow Chart

- A flowchart\* is a graphic representation of the steps that make up a process. By seeing how all of the parts of a process connect and fit together, practitioners can identify redundancies and problem areas in their work and develop plans to correct them.

Symbol	Explanation
	Use ovals to show the beginning and end points of the process.
	Use rectangles to show each process step. Describe each step in a "verb + noun" format ("Measure amount" or "Cut materials").
	Use diamonds to indicate points where a question needs to be answered or a decision must be made. Use a "yes/no" question to describe the point ("Will this delay the project?" or "Has this been approved by the corporate office?"). Make sure that all decision point answers lead to a subsequent step or loop back to a previous step in the process.
	Use inverted triangles to indicate a waiting area or build up of inventory.

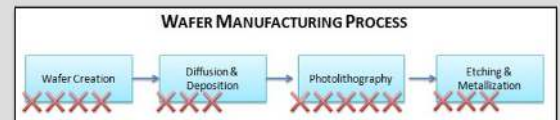
## 16. The Check Sheet

- A check sheet is a structured form or table that lets practitioners collect and record data in a simple format; by putting marks on a table or image, team members can track and record information about the number, time, and location of events or problems.
- A check sheet ensures that everyone collecting data is compiling and recording it in a similar way. Providing a structured form for recording information enhances consistency in the collection process and makes the analysis or import of the information into a subsequent evaluation process quicker and easier for all involved.

Check Sheet: Table Format

CUSTOMER SERVICE CALLS						
	DATE					
Category	5-Mar	6-Mar	7-Mar	8-Mar	9-Mar	TOTAL
Product Features						32
Product Availability						30
Product Problems						18
Cost						26
Billing						47
Other						34
TOTAL	39	33	39	35	42	

Check Sheet: Graphic Format

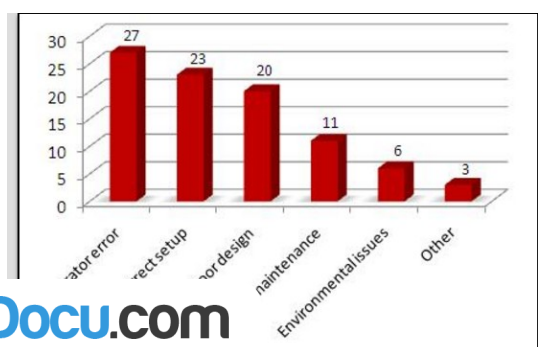


## 17. Histograms and Bar Charts

- A histogram\* is a graph that displays continuous data. The vertical bars in a histogram show the counts or numbers in each range. A comparison of the ranges, or a review of the graph as a whole, helps the audience understand the information presented.
- While a histogram measures how continuous data is distributed over various ranges, a bar chart measures data that is distributed over groups or categories.

## 18. The Pareto Chart

- A Pareto chart\* is a bar chart that sorts data into categories, then prioritizes those categories to help project teams identify the most significant factors or the biggest contributors to problems. By focusing on the factors that contribute the most to problems, practitioners can make quick and meaningful improvements to processes and work.
- The Pareto chart is based on the 80/20 rule\*, which states that 80% of quality management problems are the result of a small number (about 20%) of causes. By



concentrating your efforts to fix these "vital few" causes, you can quickly produce the greatest impact on work in the most cost-effective way possible.

- c. Focus your initial attention on the two or three largest bars in the chart; addressing the issues that these bars represent usually provides the greatest potential for improvement. If your chart does not show prominent differences among the heights of the bars, consider new categories or ways to segment the data and replot the chart.

#### 19. The Scatter Diagram

- a. A scatter diagram\* helps to show potential relationships or correlations between two variables. Data points are plotted as dots along an XY axis, and the concentration or dispersion of these dots shows the strength of the relationship between the variables.

#### 20. Decision making with the Seven Quality Tools

- a. The Ishikawa tools provide a straightforward, systematic way for project teams to address quality issues. These tools allow practitioners and team members to use a common language to understand and explore options as they address problems. This common language ensures that everyone is focused on the same issue, at the same time, in the same way so participants can compare information and discuss solutions in a way that everyone understands.
- b. The Ishikawa tools can be very helpful on a stand-alone basis to objectively:
  - i. Expose problems
  - ii. Evaluate processes
  - iii. Track progress
  - iv. See correlations and relationships among issues and results
  - v. Find the root cause of a problem
  - vi. Explore and prioritize options
  - vii. Align resource so corrective actions can be implemented
  - viii. When used in combination, the Ishikawa tools can be a powerful resource for digging even deeper into problems and for gaining an even better understanding of quality issues.
- c. When used in isolation, the seven Ishikawa tools provide a simple, objective, and powerful way to uncover quality issues and examine potential solutions. Used in conjunction, these same tools can dramatically increase the power of the team to understand the issues they face and to develop more-effective solutions to quality problems.

#### 21. "New" Tools to analyze non-numerical data

- a. These seven tools are interactive, team-based tools that help groups analyze verbal data and make better decisions on projects. Used separately or in combination, these tools help teams identify problems, generate ideas, develop plans, and implement solutions.
  - i. affinity diagram
  - ii. interrelationship digraph (or relations diagram)
  - iii. tree diagram
  - iv. process decision program chart
  - v. matrix diagram



- vi. prioritization matrix
- vii. network diagram
- b. The graphic expression of these tools allows a significant amount of information to be seen without forcing interested parties to wade through complicated tables and spreadsheets. And the tools can be linked together or used in conjunction with the Ishikawa tools to dig deeper into problems and gain a better understanding of project work.

## 22. Affinity Diagram

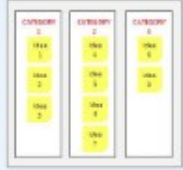



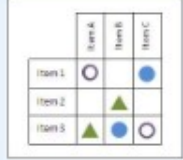
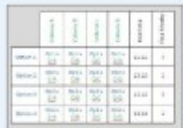
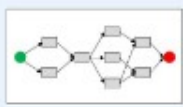
- a. An affinity diagram\* helps a team make sense of a large amount of data by grouping ideas into categories that can then be analyzed or evaluated. By sorting data into groups or categories, common themes or patterns emerge that can be used to break through bottlenecks or allow the team to clarify its thinking about issues or problems.
- b. The collaborative way that affinity diagrams are developed helps to ensure that all viewpoints are considered and prevents dominant personalities from taking over the process.

## 23. The Interrelationship Digraph

- a. An interrelationship digraph\* (or relations diagram) shows the cause-and-effect relationships among the many factors or contributors to a problem. These factors or contributors are placed in a circle and arrows are drawn to show how each factor relates to every other factor for the issue under investigation. By visualizing all of these relationships at the same time, practitioners can identify root causes and develop effective strategies for addressing multiple factors simultaneously.
- b. Interrelationship digraphs help teams visualize relationships as a network of factors rather than as a linear cause-and-effect concept. A single factor is often affected by many drivers and may itself be the driver of many subsequent factors; seeing these contributors as both inputs and outputs of other factors helps project teams identify the links among ideas and plan solutions that address the key drivers that have the greatest impact on work.

## 24. The Tree Diagram

- a. A tree diagram\* shows a hierarchy of items in a graphic format. The tool uses successive steps to break things down into greater detail, helping project teams map out their path toward a goal or showing them a logical way to work toward a result.

Tool	Description	Example
Affinity diagram	A tool that creates groups of items based on relationships which are then further investigated	
Interrelationship digraph	A circular diagram that shows the cause-and-effect relationships among contributors	
Tree diagram	A hierarchical tool that breaks a topic down into its components	
Process decision program chart	A tree diagram that illustrates options for preventing or solving problems	
Matrix diagram	A table or chart that shows the strength of relationships between items or sets of items	
Prioritization matrix	A table or chart that helps a team prioritize multiple options, based on how well these options satisfy preselected criteria	
Network diagram	A scheduling diagram that shows the relationships among project activities	

- b. Tree diagrams can be used to uncover alternatives or to divide tasks into smaller increments for assignment. The diagram is usually built from the top down, with each successive layer showing more components or options.

#### 25. The Process Decision Program Chart

- a. A process decision program chart\* (PDPC) is a tree diagram\* that is specifically designed to help teams mitigate risks and solve potential problems. The team breaks a goal or objective down into major steps or activities, then devises countermeasures or preventive actions to solve any potential problems that could happen during those activities.
- b. By anticipating problems and eliminating barriers to project success, teams can lower project costs, prevent rework, limit delays that could derail the schedule, and increase the safety and stability of project processes.

#### 26. The Matrix Diagram

- a. A matrix diagram\* shows how strong the relationships between items or sets of items are. Items from one group are compared to items in another group; the comparisons are then compiled in a table for analysis and easy reference. The results from these comparisons help to expose gaps that should be addressed and focus the team's attention on high priority areas that should be tackled immediately.
- b. It is also important to leave a significant gap between the marker for strong relationships and the marker for medium relationships, to dramatically separate these items from each other as they are ranked. If this gap is not large enough, it may be difficult to spot the differences in the data, especially if numbers are used and the columns or rows are averaged or totaled
- c. Compare each item in each row to each item in each column. If no relationship between these items exists, leave the intersecting cell blank. If a relationship does exist, place one of the markers you've chosen in the correct cell to indicate the strength of the relationship between the items.

#### 27. The Prioritization Matrix

- a. A prioritization matrix\* helps a team prioritize options when there are multiple criteria to satisfy at one time. By agreeing on the criteria that are most important to the project and then comparing options based on that information, teams can focus on implementing the best actions to quickly make an impact.
- b. Every project has criteria that it must satisfy to be considered successful but satisfying all criteria at once may not be possible, especially if resources are limited. There may also be several competing criteria that must be considered simultaneously, and each criterion may be more or less important to project success than other competing criteria.
- c. Weighting Factor is an indicator of how important a criterion is to the completion of an objective
- d. Give the team 100 points for each column and ask them to divide the points up within the column to say how well each option meets that specific criterion. (Again the points do not have to be divided equally among the options in the column.) Add this option ranking\* to the appropriate cell. Repeat this process for each column in the table, adding the option rankings to the appropriate cells. Multiply the weighting factor by the option ranking to find the weighted score\*. Record this score in the appropriate cell of the chart. Add together the weighted scores for each row to get a row total for each row. Use this row total as the final score for each option, to help you select or prioritize the options in question.

#### 28. The Network Diagram

- a. A network diagram\* is a simple graphic that shows the relationships among project activities and guides a team in executing project work. The diagram shows the predecessors (any activities that must be completed before) as well as the successors (those activities that follow) for all of the activities in the project.
- b. A network diagram helps the team:

- i. Ensure that the project will continue to meet its expected schedule
  - ii. Compare actual results to plans and see when corrections to the schedule are needed
  - iii. Keep stakeholders informed about changes to the project's time line
- c. Place the project's first activity on the far left of a large work surface. Place any activities that can be completed at the same time above or below this first activity. Next, place all subsequent activities to the right of this first activity (or activities) in a sequence that shows progress toward the project goal.
- d. To calculate the early start dates\* (ES), start at the far *left* of the diagram and determine the earliest that each activity can start, based on the completion of any predecessor activities. To determine the early finish dates\* (EF), add the amount of time an activity will take to the ES. If an activity has two predecessor activities, the earliest that the activity can start will be the latest EF of all of the predecessors. Work your way across the diagram *from the beginning of the project to the end* and calculate the ES and EF for all activities.
- e. To calculate the late finish dates\* (LF), start at the far *right* of the diagram and determine the latest that each activity can start without delaying the project. (Because the late finish date is based on the completion of any *successor* activities, the late finish date for an activity is equal to the earliest late start date of any connected succeeding task.)
- f. To calculate the late start dates\* (LS), subtract the amount of time an activity will take from the late finish dates. Work your way across the diagram *from the end of the project back toward the beginning* to calculate the LF for all activities. I
- g. The critical path\* is the path through the diagram that has no flexibility in the time allotted for path activities (i.e., any delay in completing the activities on this path will likely delay subsequent activities and the project's completion date).

## 29. Quality Management Methodologies and Programs

- a. A quality management methodology is a combination of tools, techniques, and ideas that help an organization manage its quality processes and activities. These practices embed quality concepts into the work that practitioners do every day, making sure that quality becomes an ingrained part of company culture. Some of the more popular methodologies are:
  - i. Lean - Lean practices aim to increase organizational quality by eliminating waste and reducing the time that it takes to complete processes. By eliminating all activities that do not add value or satisfy customers, lean activities focus on only the things that will help the organization achieve its objectives. And by optimizing the system--the *entire* system--lean practitioners increase their ability to produce high-caliber products and complete successful projects.
  - ii. Six Sigma - Six Sigma is a statistics-based methodology that works to eliminate variation and increase the capability of processes to meet requirements successfully. The methodology is labeled Six Sigma because its focus is the control of a process to the point of plus or minus six sigma--or standard deviations--from a centerline, or 3.4 defects per million items. The activities that Six Sigma companies execute work to reduce the number of process defects or missed opportunities to a minimal level, thereby maximizing a process' ability to deliver satisfactory results. The fundamental objective of this highly disciplined, data-driven approach is to stabilize work by driving variation out of company processes.
  - iii. Lean Six Sigma - Lean Six Sigma combines the two previously mentioned methodologies into a hybrid practice that capitalizes on the best qualities of both disciplines. By selecting specific tools from each discipline, Lean Six Sigma combines the speed of Lean with the quality control of Six Sigma. This combined approach keeps an intense focus on customers as it works to eliminate waste and simultaneously reduce variation in its processes.



- iv. Design for Six Sigma - A Design for Six Sigma approach attempts to incorporate Six Sigma ideas and methods directly into the design process. By thinking about increased capability and reduced variation as they design processes and products, teams prevent the rework and reengineering that could be needed later.
- v. ISO Certification - Organizations often pursue International Organization for Standardization or ISO certification to show that they are dedicated to the continual improvement of their processes. Achieving ISO certification ensures that an organization has put a quality management system in place to guarantee that their products satisfy the high-quality requirements they were designed to meet.

### 30. Lean Systems

- a. In an effective Lean system:
  - i. Practitioners work continuously to eliminate the defects and rework in their systems
  - ii. Material and information is pulled by downstream processes in the system
  - iii. Value-added work is enhanced and non-value-added work is minimized
  - iv. Resources (including employee skills) are maximized to the greatest extent possible
  - v. Inventory is minimized and delivered just-in-time
  - vi. Problems are quickly identified at their source and resolved as soon as possible
  - vii. Quality, productivity, performance, cost reduction, and information sharing are continuously improved
- b. Lean is focused on the "customer." It is important to remember that a customer is not always external. In the contexts of processes, a customer is whoever is receiving the outputs of your process.
- c. Lean uses concepts to expose problems and to streamline delivery throughout the *system*, rather than in *individual departments or pockets*. Lean practitioners view the problems they uncover as "opportunities" that allow them to refine and improve their work
- d. Lean organizations build reliability, adaptability, and flexibility into their processes so they can deliver superior performance for customers, employees, and stakeholders. Lean practitioners look at work from the customer's point of view in an attempt to continually uncover (and ultimately remove) any waste that they believe customers would not pay for.
- e. Lean frees up capacity within an organization to allow businesses to provide more value. By eliminating any activities that do not add value for customers and perfecting those activities that do provide the results that customers are looking for, Lean organizations free up capital and resources, which allows them to dedicate these newly liberated resources to the results required by customers.

### 31. Value Stream Mapping

- a. One of the best ways to view a lean process within a system is to look at a value stream. Value streams contain all of the steps, processes, and communication involved in the production of goods or services. A value stream shows the value-added and non-value-added work in all organizational processes (from the influx of raw materials to the delivery of finished goods to customers), making it a "systems approach" for looking at—and refining—the work that an organization needs to do.

### 32. The "Lean House"

- a. The ideas and concepts embraced by Lean practitioners are often depicted as a house with two pillars, like the one illustrated below.
- b. Just – In – Time - A plan to produce and deliver exactly what a customer needs at the precise time it is needed and in the



specific quantity needed. Just-in-time production allows organizations to continue to meet customer requirements using the fewest resources and lowest costs possible.

- i. The concept is often paraphrased as "providing customers and stakeholders\* with the *right parts* in the *right place* at the *right time*."
  - ii. The ultimate goal of just-in-time systems is continuous flow\*—moving one unit or piece at a time to the next process step or workstation as determined strictly by customer (downstream\*) demand.
- c. Flow- The continuous, smooth, and sequential movement of single pieces or very small batches of product through the value stream. Flow reduces costs (because it uses fewer resources and eliminates excess inventory) and shortens production times (because goods can be moved one-at-a-time to the next processing step instead of having to wait for an entire batch to be done).
  - d. Pull - The production of parts and products based strictly on customer or downstream demand. Pull ensures that the right parts are in the right place at the right time and prevents overproduction because products are not produced until a downstream process or customer asks for them.
  - e. Takt Time - calculation that shows how often products need to be produced to keep up with customer demand. Takt time allows organizations to have products or services ready in a just-in-time manner.
  - f. Heijunka - A coordination concept that helps to balance the type, quantity, and timing of production. Rather than scheduling production to make large batches of each part before moving on to another one, production is alternated to make smaller batches of different parts in the order and amount needed next by customers.
  - g. Jidoka- The practice of transferring intelligence and error-detection responsibilities from human workers to the machines and workstations that produce the work. Machines and workstations are able to detect abnormalities and immediately stop themselves so defects are not produced and perpetuated.
  - h. Poka-yoke A simple device that makes it physically impossible for defects or errors to be created. A poka-yoke device is configured so that parts or products must be in the correct sequence or position before processing can proceed.
  - i. Andon - A visual signal that alerts everyone that a problem exists and needs to be fixed immediately. When an andon is activated, production stops until the problem is resolved.
  - j. Gemba - The "actual place" where an activity or work occurs. Lean practitioners are encouraged to go to the gemba to see for themselves how things are actually done, rather than relying on hearsay or another person's observations or opinions.
  - k. Standard Work - A set of very specific instructions that shows the best way to complete the work of a particular workstation. Standard work specifies what steps must be taken and in what order they must be completed; this allows workers to concentrate on improvements and innovations instead of the mundane aspects of their work.
  - l. Kaizen - A continuous improvement concept that engages employees at all levels of an organization in revising and refining the value stream. Kaizen is usually comprised of many incremental activities that add up over time to create substantial improvements.

### 33. A Change in Culture

- a. All Lean practitioners need to understand that the continuous improvement of activities is everybody's responsibility and, as such, everyone must remain constantly vigilant for improvement opportunities. All parties will need to continue to search for ways to improve the quality, productivity, performance, and interaction in all aspects of their work.

### 34. Six Sigma

- a. Six Sigma\* is a highly disciplined, data-driven approach for improving quality. The Six Sigma methodology focuses intently on facts and statistics, allowing practitioners to quantitatively measure performance and progress as corrections are made.
- b. This six-sigma level is a statistical concept that places six standard deviations between the mean (or average) value for quality outputs and allowed limits on a histogram\* used to measure results. Processes working at a six-sigma level are 99.9997% defect-free (or only 3.4 defects per million process outputs). By using this high sigma level as a benchmark for reducing mistakes and systematically eliminating the defects that endanger processes, practitioners can increase process capability and achieve superior performance in their work.

### 35. The DMAIC Framework

- a. Six Sigma employs a five-step framework to analyze an existing process and to incorporate changes that will improve its ability to meet requirements:
- b. Critical To Quality characteristics\* (CTQs) that represent the attributes that internal and external customers use to evaluate the quality of process outputs. Appropriately addressing these measurable characteristics in process results will increase customer satisfaction and enhance customer loyalty.
- c. The CTQ tree is a tree diagram\* that breaks customer needs or expectations down into values that can be measured and monitored.

The customer needs (at the top of the diagram) are related to specific measurable data (at the bottom of the chart) that will show whether these needs are being met.

DMAIC Step	Explanation
D	<p>Define the project's purpose and scope, and collect the customer requirements (i.e., the <b>Voice of the Customer</b>)</p> <p>Tools: <b>control chart, Pareto chart, run chart, SIPOC diagrams</b></p>
M	<p>Measure the work's current state and uncover existing problems</p> <p>Tools: <b>control chart, flowchart, histogram, Pareto chart, run chart</b></p>
A	<p>Analyze any problems found and use data to expose root causes</p> <p>Tools: <b>cause-and-effect diagram, Design of experiments, interrelationship digraph, scatter diagram, tree diagram</b></p>
I	<p>Improve processes by implementing solutions to optimize the current state</p> <p>Tools: <b>network diagram, control chart, histogram, Pareto chart, Plan-Do-Check-Act cycle, prioritization matrix, run chart</b></p>
C	<p>Control the new processes and monitor them to maintain any advantages gained</p> <p>Tools: <b>control chart, Plan-Do-Check-Act cycle, run chart</b></p>

### 36. Reliability of a System Series

- a. System reliability is important in any instance where the success of the whole system relies on the success of each component. In other words, if any component fails, the system or process fails.

### 37. Reliability with a Parallel Process

- a. In a parallel process, if any component works, the entire system works. All components have to fail for the system to fail. To determine the reliability of the system, multiply the likelihoods of each component in the system NOT working.

## Module 5: Real World Data-Driven Decisions

### 1. Results-based Management

- a. Results-based management (RBM)\* is a management strategy that uses results as the central measurement of performance. It has been adopted by many nonprofit and governmental institutions.
- b. The United Nations Development Group more formally defines RBM as "a

RBM Stage	Explanation	Example
1. Input	The first step of RBM is to define the resources, human or financial, used by the RBM system. This could include the people who are doing the work, the funds used to finance the work, and the information being accessed.	A nonprofit identifies the people and funding needed for a special project to extend its counseling services to recent immigrants.
2. Activities	This second step involves the process that converts inputs to outputs. It includes the actions necessary to produce results. This could include training, evaluating, developing action plans, or working with media.	The nonprofit develops an action plan, recruits new counselors, and prepares a media campaign to reach recent immigrants.
3. Output	This next step is when the outputs have been created by the RBM activities. This could include goods and services, publications, systems, evaluations, or changes in skills.	The nonprofit's media campaign to reach recent immigrants launches on local radio and television stations.
4. Outcome	This is the short-term effect that the outputs will have. It could include greater efficiency, more viability, better decision making, social action, or changed public opinions.	A significant number of the nonprofit's new clients are recent immigrants seeking counseling.
5. Impact	This last step when applying results-based management is to study the long-term effects that the output will have. This could include economic, environmental, cultural, or political change.	The number of recent immigrants seeking help from the nonprofit grows as the community becomes more aware of the services being offered.

management strategy by which all actors on the ground, contributing directly or indirectly to achieving a set of development results, ensure that their processes, products and services contribute to the achievement of desired results (outputs, outcomes and goals). RBM rests on clearly defined accountability for results and requires monitoring and self-assessment of progress towards results, including reporting on performance."

- c. RBM takes a life-cycle approach\* in which the processes are continuous and cyclical.
- d. As it is a life-cycle approach, RBM requires constant planning, monitoring, evaluating, and executing the plan.
- e. Transparency, simplicity, and flexibility are also vital to a RBM system's success. All stakeholders need to have clear and well-defined roles, and they must also understand how they and the organization benefit from the realization of goals.

## 2. Big Data

- a. As mentioned earlier in this course, big data\* refers to very large amounts of data. Before the advent of the modern computer and the Internet, data was calculated in a small-scale, piecemeal fashion.
- b. Computing power available to many companies today allows us to create, maintain, and analyze huge amounts of data, although often with the help of data analysts and sophisticated data management techniques.
- c. Parsing through these numbers, analyzing the data, and making informed decisions based on the results, all components of "big data," will continue to drive managerial decision making.

## 3. Data Mining

- a. Data mining\* is the process of discovering patterns in large data sets. Data mining is performed on big data to decipher patterns from these large databases.
- b. Data mining has its challenges, which can be exacerbated with an increase in scale. Although large sample sizes are beneficial for statistical analyses, ironically one of the problems with big data is that there is too much of it. Being able to parse relevant data and differentiate between causation and correlation becomes nearly impossible as the databases become huge.
- c. Data mining will often find trends, but overlook what the underlying causes might be.

## 4. The Value of Big Data

- a. A report from the McKinsey Global Institute ("Big data: The next frontier for innovation, competition, and productivity") outlines five broad ways in which Big Data creates value. These five ways are:
  - i. Big Data makes "information transparent and usable at much higher frequency";
  - ii. It allows the collection "more accurate and detailed performance information on everything from product inventories to sick days" and therefore exposing variability and boosting performance;
  - iii. It permits the development of more tailored products or services by allowing narrow segmentation of customers;
  - iv. It facilitates analytics that can improve management decisions;
  - v. It helps the development of next generation products and services through data analysis

## 5. Business Improvement Analytics

- a. A common analytic for business improvement, index numbers\* are employed to measure the change in quantity or price over time for a good or a number of goods and services.
- b. One of the most commonly used index numbers throughout all of business is the consumer price index\*, or CPI. The CPI is a defined "basket" of assorted consumer goods and services that are purchased by a common household.
  - i. The CPI is significant because it is the main measure of inflation for a country. This number gives everyone an idea of how the economy, as a whole, has changed over time.

- c. Base Period – a period in time used as a point of reference when being compared to other time periods
    - i. The base period is important because it is used to give an initial reference point to compare all other things to.
  - d. A simple index number\* shows the change in price or quantity of a single good or service over time. This can be determined in three steps.
    - i. Determine the price or quantity of a good at the base period (first time that the researcher is comparing).
    - ii. Determine the price or quantity of a good at the time being compared to the base period.
    - iii. Consider the following variables:
      - 1.  $I_t$  = index number
      - $Y_t$  = price at time being compared to base period
      - $Y_0$  = price at base period
    - iv. Use the following equation to calculate the index number.
      - 1.  $I_t = (Y_t / Y_0) * 100$
  - e. A composite index compares the prices or quantities of a number of goods or services over time. There are two main types of composite indexes: simple and weighted.
6. Simple Composite Index
- a. A simple composite index\* is created when a researcher gathers data from many different sources without weighing any data more significantly than any other data.
7. Weighted Composite Index
- a. A weighted composite index\* is created when a researcher applies more weight to certain goods or services than others as they are calculating the index number. More weight gets applied to the certain goods or services based on quantity sizes or prices. This index gives an understanding that is more proportionate to actual changes over time.
    - i. There are different ways to weigh the different goods or services. Let's look at two of these weighted composite indexes: the Laspeyres index\* and the Paasche index\*.
    - ii. The Laspeyres index often overstates inflation, whereas the Paasche index often understates inflation.
8. Health Care Analytics
- a. Nearly every industry in the 21<sup>st</sup> century is data driven, and the healthcare industry is no exception. We have alluded to there being many applications of the concepts we've learned within the field of medicine, but in addition to these statistical concepts, there are a variety of analytics designed specifically for the healthcare industry.
9. Epidemiology
- a. Epidemiology\* is the study of the incidence, distribution, and possible control of diseases and other factors relating to health. Public health and medical professionals rely on these analytics to help shape the decisions they make every day. Rate, ratio, and proportion are fundamental elements of statistics that are essential to epidemiology.
  - b. A rate\* is the measure of an event occurring over a period of time
  - c. A ratio\* measures one quantity in relation to another quantity. For example, the gender ratio measures the number of males in a population to the number of females in the same population.
  - d. a proportion\* is a ratio in which a part of a group is compared to the whole group. For example, a proportion can be used to measure the number of people with diabetes in relation to the total population.
10. Prevalence and Incidence

- a. Prevalence and incidence are two similar measures used when studying disease frequency, but they do have differences, and it is important to understand when to use each one.
- b. Prevalence\* counts all of the existing cases of a disease, while incidence\* only counts *new* cases
- c. Prevalence measures the number of cases of a particular disease that exist in a population. This measure does not differentiate between cases — it counts both existing cases and new cases. Because it measures the number of individuals that have a specific disease compared to the population, prevalence is a type of proportion. Prevalence, P, can be calculated as follows:
  - i.  $P = (\text{number of cases})/(\text{total population})$
- d. Incidence, on the other hand, measures the number of new cases that arise in a population over the course of a designated time period. There are two types of incidence that are commonly used; cumulative incidence\* (CI) and incidence rate\* (IR).
- e. Cumulative incidence is a proportion that measures the number of new cases that arise in a particular time period compared to the population. Cumulative incidence, CI, can be calculated as follows:
  - i.  $CI = (\text{number of new cases in a particular time period})/(\text{population})$
- f. The other commonly-used type of incidence is incidence rate. A rate is a proportion that includes an amount of time in the denominator. When studying disease frequency, incidence rate uses person-time units, which is a cumulative measure of the amount of time each person was studied. Incidence rate, IR, can be calculated as follows:
  - i.  $IR = (\text{number of new cases})/(\text{person-time units})$

#### 11. Education Analytics: Data Driven Decision Making in Education

- a. ANOVA - Analysis of Variance (ANOVA) is a technique used to determine if there is sufficient evidence from sample data of three or more populations to conclude that the means of the population are not all equal. Ordinary hypothesis testing can be used to determine whether there is a basis to conclude that two populations have different means.
- b. Percentiles - A percentile signifies the percent of the population that falls below a certain value. If someone was in a class of 30 and he or she was in the 90<sup>th</sup> percentile for a test, there would be three people who had better scores.
- c. Standard Scores - Standard scores, or z-scores, measure the distance of a piece of data from the mean compared to the entire population. The unit of measure for a standard score is a standard deviation.
- d. Test Construction - There are two types of tests: norm-referenced tests and criterion-referenced tests.
  - i. Norm-referenced tests\* compare an individual to other individuals. A standard score is type of norm-referenced measurement.
  - ii. Criterion-referenced tests\* compare an individual to certain defined standards. A driver's permit test where someone has to answer a certain number of questions correctly in order to pass is an example of a criterion-referenced test.

#### 12. True Score Model – Classical Test Theory

- a. An observed score\* is the score that is actually achieved by an individual on a test.
- b. A true score\* is the average score an individual would achieve if he or she were to take the test infinite times.
- c. Because this uncertainty exists, it is important to consider measurement error. There are two types of measurement error: random measurement error\* and systematic measurement error\*.
- d. Random error is because of chance coincidental situations. Random error can eventually be eliminated over time as the number of tests increases.
- e. Systematic errors occur when something that is entirely unrelated to the test is affecting the results.

- f. The True Score Theory\* states that, in a test without systematic error, the observed score is the true score plus the random error. Observed or true scores and types of error are incredibly important when creating and analyzing a test because they help determine the validity of those tests.

### 13. Item Response Theory

- a. Item Response Theory\* (IRT), also known as latent trait theory, is a model of designing, analyzing, and scoring tests. It is often considered more useful than classical test theory because it tries to pull as much information out of each answer as possible.
- b. IRT looks at each individual question and tries to determine the meaning of a correct answer. There are a number of different IRT models, including the three parameter logistic (3PL) model, the logistic and normal IRT model, and the Rasch model. IRT can be more difficult to apply than classical test theory but generally gives stronger findings and does a better job of scaling people.

### 14. Reliability of a Test

- a. A reliability index\* is taken to determine how well a person's observed score on a test represents that person's true score. This is the ratio of the standard deviation of the true scores and the standard deviation of the observed scores.
- b. In the test-retest method\*, the same test is given to the same sample on two different occasions. Correlation is lost as the time between the tests increases.
- c. In the parallel-forms method\*, two tests are made in the exact same way and given to the same sample. The parallel-forms method and the test-retest methods compare the two samples to each other.
- d. The split-half method\* and Cronbach's Alpha\* measure internal consistency, and therefore do not take any information outside of the one sample test's results.
  - i. The split-half method splits a sample into halves and measures the correlation between the two halves.
  - ii. Cronbach's Alpha takes the average of all of the possible variations of the split-half method within a sample.

### 15. Item Analysis

- a. Item analysis\* is the study of the results for each individual item, or question, on a test. This is very important in education as it can show how a test should be adjusted and gives insight into an individual's test results.
- b. The following equation determines the likely observed score to account for this:  $p = p_0 + (1 - p_0) / m$ 
  - i. Where:
    - 1.  $P$  = item's observed score
    - ii.  $p_0$  = item's difficulty (true score)
    - iii.  $m$  = item's number of possible answers
- c. This takes into account the difficulty of an item as well as the chances that someone randomly picks the correct answer from the options available.

### 16. Public Sector Analytics

- a. Analytics is of increasing importance in the public sector. In the private sector, companies use analytics to maximize their profits and cut costs. Similarly, analytics are used in government to understand past performance and deliver public services at a lower cost.
- b. To achieve their goals and better allocate their resources, government agencies often use cost-benefit analysis. Cost-benefit analysis is seen as a staple of analytics within the private sector, but it is a technique that actually began in the public sector.
- c. In a cost-effective analysis, a quantifiable goal is determined, such as recognition of an enemy assault. To conduct the cost-effective analysis, the cost of achieving the predetermined goal is analyzed.



## 1. Performance Improvement

- a. Creating this organizational alignment around performance improvement is not easy. It requires effective leadership, communication, and visual tools to keep people engaged in the process and aware of progress updates.
- b. Performance indicators can measure aspects of financial performance, quality of programs or services, customer satisfaction, employee retention, safety statistics, energy consumption, or just about anything that can be tracked and quantified.

## 2. Performance Measures

- a. Performance measures can be used to evaluate the efficiency of an individual, a group, or even an entire organization using data collection and analytics.
- b. Critical success factors\* are those factors (quality, customer service, efficiency, etc.) critical to an organization achieving its goals.

## 3. Performance Measures and Performance Strategy

- a. Organizations can employ many different measurement frameworks or systems. These include: key performance indicators or KPI's, Kaplan and Norton's balanced scorecard, competitive benchmarking, or other approaches, including Bain's net promoter score.
- b. The Balanced Scorecard, which was developed by Robert S. Kaplan and David P. Norton, integrates four sets of measurements, financial, business process, customer, and learning and innovation, complementing traditional financial measures with those driving future performance.
- c. Benchmarking typically involves comparing an organization's performance with that of competitors or against industry standards.
- d. Net Promoter Score, developed by Fred Reichfeld, Bain and Company, and Satmetrix in the 2000s, quantifies the strength of an organization's customer relations.
- e. In broad terms, performance measures need to be: Objective, accurate, timely, easy to understand, cost effective to collect and monitor, and trackable.
- f. Performance Measurement involves assessing progress toward achieving performance goals, which include operational and financial targets. This typically involves comparing an actual performance against a baseline measurement.

## 4. Key Performance Indicators (KPIs)

- a. A key performance indicator\*, or KPI, is a performance measurement that organizations use to quantify their level of success. Choosing the appropriate performance indicator is very important, as it needs to align with the organization's definition of success.
- b. Depending on the goal of the organization, each of the following could be appropriate KPIs:
  - i. sales increase
  - ii. average order size
  - iii. employee turnover
  - iv. return customer rate
- c. KPIs often follow "SMART" criteria, which is an acronym for specific, measurable, attainable, relevant, and time-bound. Ideally, according to this acronym, a key performance indicator should select a specific area, quantify a measurable indicator, set a goal that is attainable and relevant, and state the period of time over which the measurement is occurring

## 5. KPI Performance Dashboard

- a. A performance dashboard displays key performance indicators using visual representations such as charts and graphs. If there are a number of variables that are related, a dashboard is able to bring together different graphs that represent aspects of those variables. They are used when only one chart, graph, or piece of data does not represent the complete picture.

## 6. Advantages and Disadvantages of KPIs

### a. Advantages include:

- Educate management on company performance
- Can be used as a tool across an entire organization
- Data-driven results make it easier to quantify performance
- If used over time, can create an internal benchmarking system

### b. Disadvantages include:

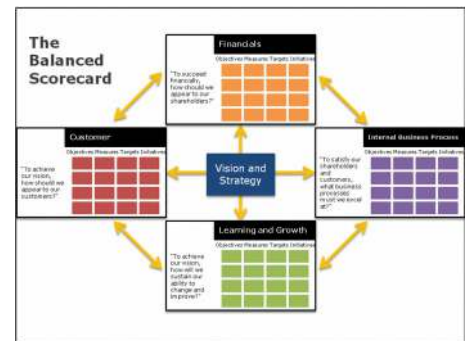
- Can be expensive and time-consuming to set up and use
- Requires frequent, even on-going, maintenance and monitoring
- Small changes in KPIs may be viewed as meaningful, but may not be statistically significant
- Results are often only rough guide rather than a concrete measurement
- Once designed and set up, difficult to change

## 7. Key Performance Indicators

- Key performance indicators, or KPI, are quantifiable measures that organizations use to gauge or compare performance against strategic and operational goals. They are also known as key success indicators, or KSI.
- Organizations will use both lagging KPIs and leading KPIs. Lagging indicators focus on the current state. For example, a lagging KPI might measure employee turnover rate. A leading indicator measures what can be done to achieve better results, for example, measuring engagement rates. In this example, the two KPIs are linked and improving employee engagement should reduce turnover.
- One benefit of using KPIs is that they can be communicated to employees, often in graphic form, and provide a clear picture of how well the organization is doing. This makes them helpful in managing daily performance. KPIs are often communicated through the use of digital dashboards, which typically display many indicators in one place.

## 8. Balanced Scorecard

- A balanced scorecard\* (BSC) measures an organization's performance on a balanced mix of both financial and non-financial measures. The purpose of the balanced scorecard to include in a company's goals some objectives that may not affect the company's current financial performance but affect the company's long-term performance.



- Financial: The financial measures on a balanced scorecard might include such items as operating income\*, revenue growth, revenue from new products, gross margin percentage, cost reductions, cash, economic value added (EVA)\*, and return on investment (ROI)\*.
  - Operating income: earnings before Interest and Taxes.
  - Economic Value Added (EVA) – net income (after taxes) earned in excess of the amount of net income required to earn the company's cost of capital
  - Return on Investment – the ratio of income earned on the investment to the investment made to earn that income.
- Customer: The customer measures on a balanced scorecard might include such items as market share, customer retention percentage, response time, delivery performance, defects, lead time, customer satisfaction\*, and the number of customer complaints.
- Internal Business Processes: The business process\* measures on a balanced scorecard might include such items as manufacturing or technological capability, new products or services, new product development times, number of new patents, defect rate, yield, average time to manufacture orders,

setup time, manufacturing downtime, time taken to repair defective products, and cycle time\*, which is the amount of time between the receipt of a customer order and the shipment of that order.

- i. Business Process – a sequence of logically related and time based work activities to provide a specific output for a customer
- ii. Cycle Time: the total elapsed time to move a unit of work from the beginning to the end of a physical process, as defined by the producer and the customer.
- e. Innovation and learning - The learning and growth measures on a balanced scorecard might include such items as employee skills, organizational learning, industry leadership, employee satisfaction scores, employee turnover rates, percentage of processes with advanced controls, percentage of employee suggestions implemented, and percentage of employee compensation based on individual and team incentives.

#### 9. Advantages and Disadvantages of Balanced Scorecard

- a. Advantages include:
  - i. Improves organization alignment
  - ii. Improves internal and external communication
  - iii. Links company operations with its strategy
  - iv. Emphasizes strategy and organizational results
- b. Disadvantages include:
  - i. Requires time and effort to establish a meaningful scorecard
  - ii. Does not illustrate a full picture of the company performance, particularly financial data
  - iii. Sometimes difficult to maintain momentum
  - iv. Requires a wide cross-section of the organization departments in developing the system
  - v. May not encourage desired behavior changes

#### 10. Organizational Performance with Balanced Scorecards

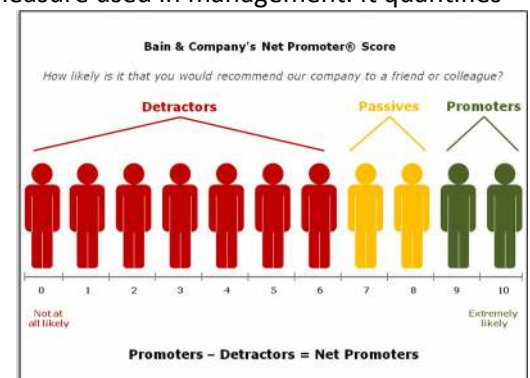
- a. The BSC has evolved over time with many practitioners advocating a tighter linkage between the scorecard and an organization's strategy map\* (a diagram documenting the primary strategic goals being pursued by an organization.)

#### 11. Effective Use of a Balanced Scorecard

- a. A successful balanced scorecard should reflect the company's strategy\* as outlined by top management. While a successful balanced scorecard should place strong emphasis on financial goals, it also should consider important non-financial measures.
- b. The number of performance measures should be limited, focusing on the most crucial objectives
- c. A calculated variance will explain if you have failed to meet expectations or if you have exceeded expectations.
  - i.  $\text{Variance} = \text{actual performance} - \text{target performance}$

#### 12. Net Promoter Score

- a. A Net Promoter Score\*, or NPS®, is a specific performance measure used in management. It quantifies how strong an organization's customer relations are.
- b. The NPS premise categorizes a company's customers into three groups:
  - i. Promoters – Enthusiastic and loyal consumers
  - ii. Passives – Unenthusiastic and satisfied but apathetic
  - iii. Detractors – Unhappy customers



#### 13. Advantages and Disadvantages of NPS

- a. Advantages include:

- i. Creates an easily understood metric for customer perceptions
  - ii. Holds employees accountable for their treatment of individual customers
  - iii. Allows companies to benchmark against industry leaders
  - iv. Relatively low maintenance if deployed electronically
- b. Disadvantages Include:
  - i. Does not provide in-depth customer perception data
  - ii. Requires customers who are willing to respond to the question
  - iii. Some argue the 11-point scale is not as predictive as 7-point scale
  - iv. Some argue it fails to predict loyalty behaviors
  - v. Can be difficult to capture the precise area of dissatisfaction

#### 14. Performance Assessment and Strategy

- a. Performance assessment can and should be linked to a company's strategy. A strategic plan gives a company a target of where it needs to be, or hopes to be, at some point in the future.
- b. The first questions that need to be asked are broad in nature: Does the organization have a mission or vision? Is it reflected in strategic plan? Are the organization's strategic goals clearly stated? Do people within the organization understand how their activities contribute to achieving the strategic goals?
- c. The next set of questions focus on the link between measurement and strategy: Are performance measures related to the mission and goals as reflected in the strategic plan? Are the measures of success focused on outcomes? Are appropriate measures reported to individuals at different levels of the organization?
- d.



the