WEEK – 1: 26/06/23

BASIC TERMS OF STATISTICS:

1. Population : All the members of a group about which you want to draw a conclusion.
2. Sample: The part of the population selected for analysis.
3. Parameter: A numerical measure that describes a characteristic of a population.
4. Statistic: A numerical measure that describes a characteristic of a sample.
5. Variable: A characteristic of an item or an individual that will be analyzed using statistics.

BRANCHES OF STATISTICS:

DESCRIPTIVE STATISTICS:

The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

Example: The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory’s production line.

INFERENTIAL STATISTICS:

The branch of statistics that analyzes sample data to draw conclusions about a population.

Example: A survey that sampled 2,001 full- or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (AARP), discovered that 70% of those polled planned to work past the traditional mid60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.

SOURCES OF DATA:

PUBLISHED SOURCES : Data available in print or in electronic form, including data found on Internet Web sites. Primary data sources are those published by the individual or group that collected the data. Secondary data sources are those compiled from primary sources.

Many U.S. federal agencies, including the Census Bureau, publish primary data sources that are available at the Web site www.fedstats.gov. Business news sections of daily newspapers commonly publish secondary source data compiled by business organizations and government agencies

EXPERIMENTS : T A process that studies the effect on a variable of varying the value(s) of another variable or variables, while keeping all other things equal. A typical experiment contains both a treatment group and a control group. The treatment group consists of those individuals or things that receive the treatment(s) being studied. The control group consists of those individuals or things that do not receive the treatment(s) being studied.

SURVEYS : A process that uses questionnaires or similar means to gather values for the responses from a set of participants.

SAMPLING CONCEPTS:

SAMPLING : The process by which members of a population are selected for a sample

PROBABILITY SMAPLING : T A sampling process that takes into consideration the chance of occurrence of each item being selected. Probability sampling increases your chances that the sample will be representative of the population.

SIMPLE RANDOM SAMPLING : The probability sampling process in which every individual or item from a population has the same chance of selection as every other individual or item. Every possible sample of a certain size has the same chance of being selected as every other sample that has that size.

SAMPLE SELECTION METHODS :

SAMPLING WITH REPLACEMENT : A sampling method in which each selected item is returned to the frame from which it was selected so that it has the same probability of being selected again.

SAMPLING WITHOUT REPLCAEMENT : A sampling method in which each selected item is not returned to the frame from which it was selected. Using this technique, an item can be selected no more than one time.

THE BAR GRAPH:

A chart containing rectangles (“bars”) in which the length of each bar represents the count, amount, or percentage of responses of one category.

PIE CHART:

A circle chart in which wedge-shaped areas—pie slices—represent the count, amount, or percentage of each category and the entire circle (“pie”) represents the total.

THE PARETO DIAGRAM:

A special type of bar chart in which the counts, amounts, or percentages of each category are presented in descending order left to right, along with a superimposed plotted line that represents a running cumulative percentage.

TWO WAY CROSS CLASSIFICATION TABLE :

A multicolumn table that presents the count or percentage of responses to two categorical variables. In two-way tables, the categories of one of the variables form the rows of the table, while the categories of the second variable form the columns. Cross-classification tables are also known as cross-tabulation tables.

HISTOGRAM : A special bar chart for grouped numerical data in which the frequencies or percentages of each group of numerical data are represented as individual bars on the vertical Y-axis and the variable is plotted on the horizontal X-axis. In a histogram, there are no gaps between adjacent bars as there would be in a bar chart of categorical data.

THE DOT SCALE DIAGRAM :

A chart in which each response is represented as a dot above a horizontal line that extends through the range of all values. Should two or more response values be identical, the dots for these responses are stacked (placed vertically) above each other.

THE TIME-SERIES PLOT:

A chart in which each point represents the value of a numerical variable at a specific time. By convention, the X-axis (the horizontal axis) always represents units of time, and the Y-axis (the vertical axis) always represents units of the variable.

THE SCATTER PLOT:

A chart that plots the values of two numerical variables for each response. In a scatter plot, the X-axis (the horizontal axis) always represents units of one variable, and the Y-axis (the vertical axis) always represents units of the second variable.

**Descriptive Statistics For Numerical Variables**

The Mean :

A number equal to the sum of the data values for a variable, divided by the number of data values that were summed.

Many sports statistics (including baseball batting averages and football yards per reception), average SAT score for incoming freshmen at a college, average age of the workers in a company, average waiting times at a bank.

The mean represents a “balance point” in a set of data values, similar to a fulcrum on a seesaw. As the only measure of central tendency that uses all the data values in a sample or population, the mean has one great weakness: individual extreme values can distort the balance point.

The Median :

The middle value in a set of data values for a variable when the data values have been ordered from lowest to highest value. When the number of data values to be summarized is even, you perform a special calculation to determine the median because data sets with an even number of values have no natural middle value.

Many economic statistics such as median household income for a region; many marketing statistics such as the median age for buying a consumer product; in education, the established middle point for many standardized tests.

The median splits the set of ranked data values into equal-in-numbers parts. Extreme values do not affect the median, making the median a good alternative to the mean when such values occur. When the number of data values to be summarized is even, the median is calculated by taking the mean of the two values closest to the middle, when all values are ranked from lowest to highest. For example, if there were 6 ranked values, you would calculate the mean of the third and fourth values; and if there were 10 ranked values, you would calculate the mean of the fifth and sixth values.

The Mode :

The value (or values) in a set of data values for a variable that appears most frequently.

The most common score on an exam, the most likely income, the commuting time that occurs most often.

Similar to the median, extreme values do not affect the mode; unlike the median, however, the mode can vary much more from sample to sample than the median (or mean). Some sets of data values have no mode—all the unique values appear the same number of times. Other sets of data values can have more than one mode, such as the get-ready times on page 38 in which two modes occur, 39 minutes and 44 minutes, because each of these values appears twice and all other values appear once.

Quartiles :

The three values that split a set of ranked data values for a variable into four equal parts—quarters, or quartiles. The first quartile, Q1, is the value such that 25.0% of the ranked data values are smaller and 75.0% are larger. The second quartile, Q2, is another name for the median, which, as discussed on page 40, splits the ranked values into two equal parts. The third quartile, Q3, is the value such that 75.0% of the ranked values are smaller and 25.0% are larger.

Quartiles help summarize large sets of data values by allowing you to identify the 25th, 50th, and 75th percentiles. If you scored in 3.1 MEASURES OF CENTRAL TENDENCY 41 interested in mathequation blackboard (optional) Using the n symbol previously defined on page 39, you can define the median as: Median ranked value = n + th 1 2 the third quartile on a standardized test, your score was in the top 25% of all scores. If your score was equal to the third quartile, the 75th percentile, then 25% of all scores were higher and 75% were lower. If you did exceptionally well, and learned that that your score was reported as the 99th percentile, you would know that your score was in the top 1% of all scores (and therefore greater than 99% of all scores).

To quickly determine the first quartile, add 1 to the number of data values and divide that sum by 4. For example, for 11 values, add 1 to 11 to get 12 and divide 12 by 4 to get 3 and determine that the first quartile is the third ranked value. To quickly determine the third quartile, add 1 to the number of data values and divide that sum by 4 and then multiply the quotient by 3. For the same example, you would multiply the quotient 3 by 3 to get 9 and determine that the third quartile is the ninth ranked value). When the result of this arithmetic is not a whole number, select the ranked positions immediately below and above the number calculated. For example, for 10 values, the result (for the first quartile) would be 2.75 (10 + 1 is 11, 11/4 is 2.75), and you would select the second and third ranked values. With these values, do the following: 1. Multiply the larger ranked value by the decimal fraction of the original result (0.75 in the example.) 2. Multiply the smaller ranked value by 1 minus the decimal fraction of the original result (0.25 for the example, because 1 – 0.75 is 0.25). 3. Add the two products to determine the quartile value. Special case: Should the two ranked values selected be the same number, then the quartile is that number and you can skip the previous two multiplications and one addition.

The Range :

The difference between the largest and smallest data values in a set of values for a variable.

In most everyday examples, the largest and smallest values are presented and the number that represents their difference is not shown: daily high and low temperatures, stock market 52-week high and low closing prices, best and worst times for timed sporting events.

The Variance and Standard Deviation :

Two related numbers that each individually measures how a set of data values for a variable fluctuate around the mean of that variable. The numbers are related because one of them, the standard deviation, is the positive square root of the other (the variance). EXAMPLE The variance among SAT scores for incoming freshmen at a college, the standard deviation in the age of the workers in a company. The variance and standard deviation always appear as “variance” and “standard deviation” and should be accompanied by the mean.

The variance and standard deviation help you to know how a set of data values distributes around its mean. For almost all sets of data values, the majority of the values lie within an interval of plus and minus one standard deviation above and below the mean. Therefore, determining the mean and the standard deviation usually helps you define the range in which the majority of the data values occur. The simplest measure of variation might take the difference between each value and the mean and sum these differences. However, by the properties of arithmetic and the definition of mean, the result of such calculations would be zero for every set of data values—not very helpful in comparing one set to another.

**DATA ARCHITECTURE**

## **What is Data Architecture?**

[Data architecture](https://www.simplilearn.com/data-architect-article) is the foundation of an effective data strategy. According to data architecture definition, it is a framework of models, policies, rules and standards that an organization uses to manage data and its flow through the organization. Within a company, everyone wants data to be easily accessible, to be cleaned up well, and to be updated regularly. Successful data architecture standardizes the processes to capture, store, transform and deliver usable data to people who need it. It identifies the business users who will consume the data and their varying requirements.

A good approach to data architecture is to make it flow from data consumers to data sources, not the other way. The goal is to transform business requirements into data and system requirements. Companies need to have a centralized data architecture that aligns with business processes and provides clarity about all aspects of data. The individual components of data architecture are the outcomes, activities, and behaviors.

 A data architect builds, optimizes, and maintains conceptual and logical [database](https://www.simplilearn.com/what-is-database-management-article) models. They determine how to source data that can propel the business forward and how that can be distributed to provide valuable insights to decision-makers.

## **Data Architecture Principles :**

 These principles form the foundation of the data architecture framework and help build effective data strategies and data-driven decisions.

### **Validate All Data at Point of Entry**

It's important to improve the overall health of organizational data by eliminating bad data and common data errors. Design your data architecture to flag and correct errors as soon as possible. A Data Integration Platform can help do that – validate data automatically at the point of entry. This will also help minimize the time taken to cleanse and prep data.

### **Strive for Consistency**

Using a Common Vocabulary for data architecture will help users on the same project to collaborate. Shared data assets like product catalogs, fiscal calendar dimensions, etc. must use common vocabulary regardless of the application or business function. Users of such shared data must work from the same core definitions to maintain control of data architecture and data governance.

### **Everything Should be Documented**

Get into the habit of documenting all parts of your data process so that data visibility and data remain standardized across the organization. Documentation should help you keep a tab on how much data is collected, which datasets are aligned, and which applications need to be updated. Consistent documentation should work seamlessly with data integration.

### **Avoid Data Duplication and Movement**

Every time data is moved, it impacts cost, accuracy, and time. Modern data architectures should reduce the need for additional data movement to reduce cost, improve data freshness and optimize data agility.  Modern data architecture views data as a shared asset and does not allow departmental data silos. This makes it simpler to universally update data, and everyone can operate from a single version of the data.

### **Users Need Adequate Access to Data**

Data architecture books state that users must be provided the right interfaces to consume data using designated tools.

### **Security and Access Controls Are Essential**

The emergence of data security projects has made it easier to ensure unified data security. Data architectures must be designed for security without compromising access controls on the raw data.

**DATA COLLECTION**

**Data collection** is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. The data collection component of research is common to all fields of study including physical and social sciences, humanities, business, etc. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same.

**The importance of ensuring accurate and appropriate data collection**Regardless of the field of study or preference for defining data (quantitative, qualitative), accurate data collection is essential to maintaining the integrity of research. Both the selection of appropriate data collection instruments (existing, modified, or newly developed) and clearly delineated instructions for their correct use reduce the likelihood of errors occurring.

**DATA EXTRACTION & ETL**

Data extraction is the process of collecting or retrieving disparate types of data from a variety of sources, many of which may be poorly organized or completely unstructured. Data Extraction makes it possible to consolidate, process, and refine data so that it can be stored in a centralized location in order to be transformed. These locations may be on-site, cloud-based, or a hybrid of the two.

1.Extraction : Data is taken from one or more sources or systems. The extraction locates and identifies relevant data, then prepares it for processing or transformation. Extraction allows many different kinds of data to be combined and ultimately mined for business intelligence.

2. Transformation : Once the data has been successfully extracted, it is ready to be refined. During the transformation phase, data is sorted, organized, and cleansed. For example, duplicate entries will be deleted, missing values removed or enriched, and audits will be performed to produce data that is reliable, consistent, and usable.

3. Loading : The transformed, data is then delivered to a single, unified target location for storage and analysis.

**DATA STRUCTURING**

Data structuring techniques, in essence, have to do with a system where seemingly random, unstructured data can be taken as input and a number of operations executed on it linearly or non-linearly. These operations are meant to analyze the nature of the data and its importance in the larger scheme of things.

The system then divides the data into broad categories of information as found by the results of the analysis, and either stores them or sends them on for extra analysis. This extra analysis can be used to break down the data into further sub-categories or nested category trees. During the analysis, some of the data might also be found to be useless and eventually discarded.

The result of this process is structured, meaningful data that can be further analyzed or used directly to gain business insight. The journey from unstructured data to business insight is what the data structuring and processing cycle are all about. Its success often determines the success of the role of data in a particular organization.

A data structure is essentially a place where data can be stored in a structured form. Right from very basic structures like arrays which are commonly used in programming languages, data structures can nowadays take complex and intricate forms, and such are the forms that are usually called upon to work with Big Data.

**DATA FORMATTING**

Data formatting-software organization in keeping with the preset requirements (usually for computer processing) data format, formatting.

* Computer, computer machinery, operating machine, data processor, electronic computer, information retrieval device-automated calculation machine.
* Data formatting is dependent on your data purpose, data elements, and much more.
* Data formatting makes our worksheet more visual.
* You can add color, modify the font, apply styles, etc. when formatting info. Font Style: The font types are different.

**Insights on Superstore India**

**Executive Summary Dashboard**

* After plotting the data using a Symbol Map and a Bar chart, it can be easily said that the state with the highest number of sales is Uttar Pradesh as it is the largest state in India.
* The state with the least number of sales is Sikkim with the number 56.
* Daman and Diu, Dadra and Nagar Haveli didn’t have any sales whatsoever.
* From the Pie Chart in sheet no. 3, it can be clearly observed that the Eastern region has the highest profit and also has the highest number of sales. Whereas the North Eastern region as the least profit with lowest number of sales in the country.
* As observed from the sheets 4&5, we get the clear insights that Uttar Pradesh has the highest quantity of goods and hence has the highest number of sales.
* Same goes for the state with the lowest quantity of goods and the lowest number of sales, i.e Sikkim.

**Product Analysis Dashboard**

* **As observed from the first sheet, Telephones and Communication is the sub-category making the highest profit in the country.**
* **It can also be seen from sheet number 2 that Telephones and Communication is the sub-category with the highest number of sales as well. Hence we can clearly conclude that the trend : Number of Sales** ∝ **Profit.**
* **However the sub-category Tables shows a difference in this trend.**
* **Tables is the sub-category with the 3rd highest number of sales. But when it comes to profits, it is last in the whole lot.**