
Understanding Data

Data, Data Types and Data Related Terms

HIGHLIGHTS

Data and Data Analysis Concepts

Making Sense from Data: Descriptive Statistics

Statistics and Data at a Glance: What you need to know

Current Developments in Data Analysis

Preparing Data for Analysis

Data Cleansing, Data Transformation

Data Warehouse

Data, Data Types, and Data Quality

Data Types, Data Collection

Cross-sectional data, Concept of variables, discrete, continuous, and Time-series Data

Describing Data using Levels of Measurement

Types of Measurement Scale

Nominal, Ordinal, Interval, and Ratio Scales

Data Collection, Presentation and Analysis

How Data are Collected: Sources of Data for Research and Analysis

Web as a Major Source of Data

Data Related Terms Applied to Analytics

Big Data, Data mining, Data warehouse, Structured vs. Unstructured

Data, Data quality

Making Sense from Data: Data and Data Analysis Concepts

Statistics and Data at a Glance: What you need to know

Statistics is the science and art of making decisions using data. It is often called the science of data and is about analyzing and drawing meaningful conclusions from the data. Almost every field uses data and statistics to learn about systems and their processes. In fields such as business, research, health care, and engineering, a vast amount of raw data is collected and warehoused rapidly; this data must be analyzed to be meaningful. In this chapter, we will look into different types of data. It is important to note that data are not always numbers; they can be in the form of pictures, voice or audio, and categories. We will briefly explore how to make efficient decisions from data. Statistical tools will aid in gaining skills such as (i) collecting, describing, analyzing and interpreting data for intelligent decision-making, (ii) realizing that variation is an integral part of data, (iii) understanding the nature and pattern of variability of a phenomenon in the data, and (iv) being able to measure reliability of the population parameters from which the sample data are collected to draw valid inferences.

The applications of statistics can be found in a majority of issues that concern everyday life. Examples include surveys related to consumer opinions, marketing studies, and economic and political polls.

Current Developments in Data Analysis

Because of the advancement in technology, it is now possible to collect massive amounts of data. Lots of data, such as web data, e-commerce, purchase transactions at retail stores, and bank and credit card transaction data, among others, is collected and warehoused by businesses. There has been an increasing amount of pressure on businesses to provide high-quality products and services to improve their market share in this highly competitive market. Not only is it critical for businesses to meet and exceed customer needs and requirements, but it is also important for businesses to process and analyze a large amount of data efficiently in order to seek hidden patterns in the data. The processing and analysis of large data sets comes under the emerging field known as **big data, data mining, and analytics**.

To process these massive amounts of data, data analytics and mining use statistical techniques and algorithms to extract non-trivial, implicit, previously unknown, and potentially useful patterns. Because applications of data mining tools are growing, there will be more demand for professionals trained in data mining. The knowledge discovered from this data to make intelligent, data-driven decisions is referred to as **business intelligence** and **business analytics**. **Business intelligence (BI)** is a hot topic in business and leadership circles today as it uses a set of techniques and processes which aid in fact-based decision-making.

Preparing Data for Analysis

In statistical applications, data analysis may be viewed as the application of **descriptive statistics (descriptive analytics)**, **data visualization**, **exploratory data analysis (EDA)**, and predictive and prescriptive analytics. Before data can be analyzed, data preparation is important. Since the data are collected and obtained from different sources, a number of steps are necessary to ensure

data quality. These include *data cleaning* or *data cleansing*, *data transformation*, *modeling*, *data warehousing*, and maintaining *data quality*. We provide an overview of these here.

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and identifying incomplete, incorrect, inaccurate or irrelevant data and then replacing, modifying, or deleting the corrupt data. ^[1]

A **scripting** or **script language** is a programming language that supports **scripts**. Sometimes a **script** or **scripting** programs are used to automate or execute the tasks that could alternatively be executed one-by-one by a human operator. Scripting languages have applications in automating software applications, web pages in a web browser, operating systems (OS), embedded systems, and games. After cleansing, a data set should be consistent with other similar data sets in the system and suitable for further analysis. [https://en.wikipedia.org/wiki/Data_cleansing]

Data transformation: *Data transformation* is the process of converting data from one format or structure into another format or structure. It is a fundamental aspect of most data integration and data management tasks such as **data warehousing** and **data integration**. [https://en.wikipedia.org/wiki/Data_transformation].

Data warehouse: A **data warehouse** (DW or DWH), or **enterprise data warehouse** (EDW), is a system for storing, reporting, and analysis of huge amounts of data. The purpose of a DW is to create reports and perform analytics, which are core components of *Business Intelligence*. DWs are central repositories used to store and integrate current and historical data from one or many sources. The data are used for creating analytical and visual reports throughout the enterprise. The stored data in the warehouse are used for creating reports and performing analytics for the different operations and applications in an enterprise, including sales, finance, marketing, engineering, and others. Before performing analyses on the data, cleansing, transformation, and data quality are critical issues.

It is also important to note that data may be both structured and unstructured. The distinction is explained here.

Structured vs. Unstructured Data: refer to the ‘Volume’ and ‘Variety’ – the ‘V’s of Big Data. Structured data is data that can be stored in relational databases. This type of data can be analyzed and organized in such a way that it can be related to other data via tables. Unstructured data cannot be directly put in databases or analyzed or organized directly. Some examples are email/text messages, social media posts and recorded human speech, etc.

Data and Data Quality

In data analysis and analytics, data can be viewed as *information*. Data are also *measurements*. The purpose of data analysis is to make sense of data. Data, when collected (in its raw form), are known as *raw data*. These are the data not processed.

In data analysis, data needs to be converted into a form suitable for reporting and analytics.

[<http://searchdatamanagement.techtarget.com/definition/data>]. It is acceptable for data to be used as a singular subject or a plural subject. Data when collected, is often referred to as raw data, which is unprocessed data. Raw data is a term used to describe data in its most basic digital format.

Data Quality

Data quality is crucial to the reliability and success of Business Analytics (BA) and Business Intelligence (BI) programs. Both analytics and business intelligence are data-driven programs. Analytics is about analyzing data to drive business decisions, whereas BI is about reporting.

Data quality is affected by the way data is collected, entered in the system, stored, and managed. Efficient and accurate storage (data warehouse), cleansing, and data transformation are critical for assuring data quality. The process of verifying the reliability and effectiveness of data is sometimes referred to as Data quality assurance (DQA). The effectiveness, reliability, and success of business analytics (BA) and business intelligence (BI) depend on acceptable data quality.

The following are important considerations in assuring data quality Aspects of data quality include: [<http://searchdatamanagement.techtarget.com/definition/data-quality>]

- (a) Accuracy (b) Completeness (c) Up-to-date status (d) Relevance
- (e) Consistency across data sources (f) Reliability (g) Appropriate presentation
- (h) Accessibility

Within an organization, acceptable data quality is crucial to operational and transactional processes. Maintaining data quality requires going through the data periodically and scrubbing it. Typically, this involves updating, standardizing, and de-duplicating records to create a single view of the data, even if it is stored in multiple disparate systems. There are many vendor applications on the market to make this job easier. [<http://searchdatamanagement.techtarget.com/definition/data-quality>].

Much of the data analysis and statistical techniques discussed here are prerequisites to fully understanding data mining and business analytics. Besides statistics and data analysis, this book deals with several of the tools used in **Data Science**. Since statistics is at the core of data science and involves data analysis and modeling, we provide an introduction to statistics as a field before discussing applications.

Data and Classification of Data

Data are any number of related observations. We collect data to draw conclusions or to make decisions. Data often provide the *basis* for decision-making.

A single data point or observation is known as a ***data point***. A collection of data is a ***data set***. In statistics, reference to data means a collection of data or a data set.

Data can also be ***qualitative*** or ***quantitative***. ***Quantitative data*** are numerical data that can be expressed in numbers. For example, data collected on temperature, sales and demand, length, height, and volume are all examples of quantitative data.

Qualitative data are data for which the measurement scale is categorical. Qualitative data are also known as ***categorical data***. Examples of qualitative data include the color of your car, response to a yes/no question, or the product rating using a Likert scale of 1 to 5 where the numbers correspond to a category (excellent, good, etc.).

Data can also be classified as ***time series data*** or ***cross-sectional data***. ***Time series*** data are the data recorded over time; for example, weekly sales, monthly demand for a product, or the number of orders received by an online shopping department of a department store.

Cross-sectional data are the values observed at the same point in time. For example, the closing value of the stock market on the 5th of each month for the past twelve months would be considered cross-sectional because all observations correspond to one point in time.

Statistical techniques are more suited to quantitative data. These techniques involve principles and methods used in collecting and analyzing data.

Data Elements are the entities—the specific items—that we are collecting data about. For example, data collected on the stock price of companies. Each company's stock price is an ***element***.

Variable: In statistics, a variable can be thought of as an object upon which the data are collected. This object can be a person, entity, thing, or an event. The stock price of the companies above is a ***variable***. Note that stock values show variation. In the case of stock prices for the above companies, we can say that the stock price is a variable because the prices vary. If data are collected on daily temperature for a month, the temperature will show variation. Thus, the temperature is a ***variable***. Similarly, data collected on sales, profit, the number of customers served by a bank, the diameter of a shaft produced by a manufacturing company, and the number of housing starts all show variation; therefore, these are ***variables***. Using statistics, we can study the variation. A variable is generally a characteristic of interest in the data. A data set may contain one or more variables of interest. For example, we showed the stock price of five companies in the above example. There is only one variable in the data: the stock price. If we also collected data on earnings and P/E (price-to-earnings) ratio of each company, we would have three variables.

Another Classification of Data

Data are also classified as:

- **Discrete**, or
- **Continuous**

Discrete data are the result of a counting process. These are expressed as whole numbers or integers. Some examples of discrete data are cars sold by Toyota in the last quarter, the number of houses sold last year, or the number of defective parts produced by a company. All these are expressed in whole numbers and are examples of discrete data.

Continuous data can take any value within a given range. These are measured on a continuum or a scale that can be divided infinitely. More powerful statistical tools are available to deal with continuous data as compared to discrete data; therefore, continuous data are preferred wherever possible. Some examples of continuous data include measurements of length, height, diameter, temperature, stock value, sales, etc. Discrete and continuous data may also be referred as **discrete variable** and **continuous variables**.

Data Types and Data Collection

Data are often collected on a variable of interest. For example, we may collect data on the stock value of a particular technology stock, number of jobs created in a month, or diameters of a shaft manufactured by a manufacturing company. In all these cases, the data will vary; for example, the diameter measurement will vary from shaft to shaft. Data can also be collected using a survey where a questionnaire is designed for data collection purposes. The response in a survey generally varies from person to person. In other words, the response obtained is **random**.

In the data collection process, the response or the measurements may be **qualitative**, **quantitative**, **discrete** or **continuous**. If the data are quantitative, they can be either discrete or continuous. The quantitative data are also known as **numeric data**. Thus, the data can be classified as:

- **Qualitative data**, and
- **Quantitative data**

This classification of data is shown in Figure 3.1.

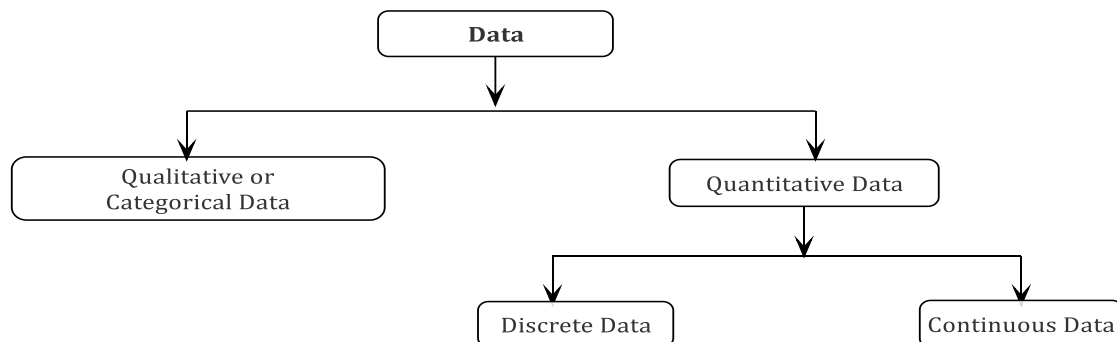


Figure 3.1: Classification of Quantitative Data

Describing Data using Levels of Measurement

Data are also described according to the **levels of measurement** used. In the broadest sense, all collected data are **measured** in some form. Even discrete quantitative data are nothing but **measurement through counting**. The type of data, and the way the data are measured, makes

them weak or strong for analysis. The data measured at different levels and are classified into different measurement scales. These are discussed here.

Types of Measurement Scales:

There are four levels or scales of measurements:

1. **Nominal Scale**
2. **Ordinal Scale**
3. **Interval Scale, and**
4. **Ratio scale**

The nominal is the weakest and ratio is the strongest form of measurement.

Nominal and Ordinal Scales

Data obtained from a qualitative variable are measured on a *nominal scale* or *ordinal scale*. If the observed data are classified into various distinct categories in which *no ordering* is implied, a *nominal level* of measurement is achieved. See the examples below.

<i>Qualitative variable</i>	<i>Category</i>		
Marital status	Married	Single	
Stock ownership	Yes	No	
Political party affiliation	Democrat	Republican	Independent

If the observed data are classified into distinct categories in which ordering is implied, an *ordinal level* of measurement is obtained.

<i>Qualitative variable</i>	<i>Ordered categories</i>
Student grades	A B C D F
Rank of employees	Senior engineer, Engineer, and Engineer trainee
Product quality	Excellent, good, poor (highest to lowest)

Nominal scale is the weakest form of measurement. Ordinal scale is also a weak form of measurement because no meaningful numerical statements can be made about the different categories. For example: the ordinal scale only tells which category is greater, but does not tell how much greater.

Interval and Ratio Scales

Interval Scale: These measurements are made on a quantitative scale. It is an ordered scale in which the difference between any two measurements is a meaningful quantity. For example, a person who is 70 inches tall is 2 inches taller than someone who is 68 inches tall. The difference of 2 inches would also be obtained if the heights of two persons were 78 and 76 inches. The difference has the same meaning anywhere on the scale. Some examples are:

<i>Quantitative variable</i>	<i>Levels of measurement</i>
Temperature	Interval

Time

Interval

Ratio Scale: If, in addition to the difference being meaningful and equal at all points on a scale, there is also a “true zero” point in which the ratio of measurements is sensible to consider, then the scale is a ratio scale. The measurements are made from the same reference point. For example, a measurement of 80 inches in length is twice as long as a measurement of 40 inches. Measurements of length are on a ratio scale. Examples of ratio scale are:

<i>Quantitative variable</i>	<i>Level of measurements</i>
Height (in feet, inches)	Ratio
Weight (in pounds, or kilograms)	Ratio
Age (in years, days)	Ratio
Salary (in dollars)	Ratio

Data obtained from a quantitative variable are measured on an interval or a ratio scale. Ratio is the highest level of measurement. It tells which observed value is largest, and by how much. Figure 3.2 shows the classification of data according to the levels of measurement.

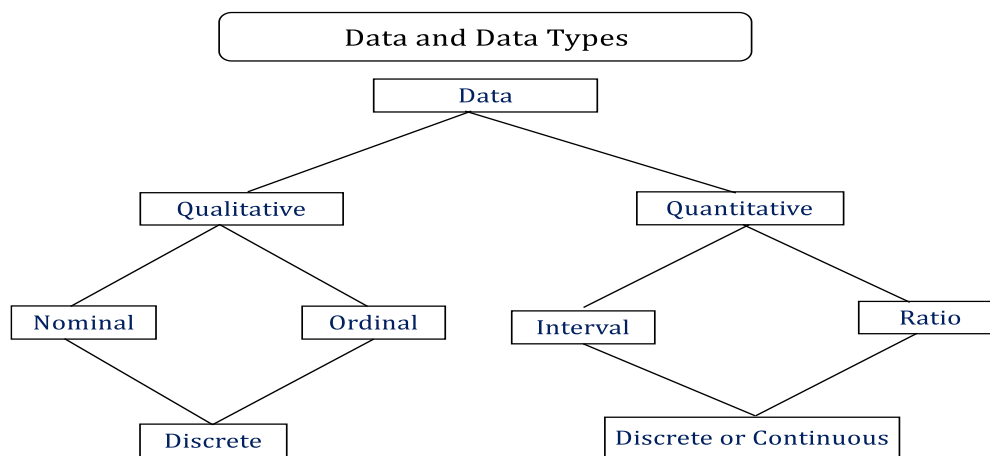


Figure 3.2 Classifications of Data

Data Collection, Presentation and Analysis

In this section we describe:

- How data are collected (obtained) and prepared for statistical analysis
- Tabular presentation of data
- Graphical presentation of data
- Analysis and interpretation of data

In statistical analysis, we always encounter situations where we need to draw conclusion from the data. Effective decision-making requires information. For the statistician or researcher, the information needed to make effective decisions is in the data. For a statistical analysis to be useful, and to make an effective decision, the input data must be appropriate. If the data are insufficient,

flawed, ambiguous, or have errors, even the most sophisticated statistical tools will not provide meaningful results.

How Data is Collected: Sources of Data for Research and Analysis

Web as a Major Source of Data

Data can be obtained from industrial, individual, or government sources. The Internet is now one of the major sources of data. A number of websites can be accessed using search engines like Google to obtain data. There are several websites that collect data on employment, consumer price index, population, housing, and manufacturing. Law enforcement agencies collect data on crimes. A great deal of business and economic data can be obtained from the web sites of Bureau of Economic Analysis (BEA), the Bureau of Labor Statistics (BLS), the U.S. Census Bureau, the Federal Reserve Economic Data, (FRED), and the National Association of Manufacturers (NAM). Besides these, a vast amount of data on current issues, stock markets, economics, sports and other areas of interest can be obtained from the websites of reputable publications including *USA Today*, *The Wall Street Journal*, *The Economist*, *Money Magazine*, *The New York Times*, to name a few. Table 3.2 provides a list of some Internet sites and data that can be obtained from these sites.

Businesses collect massive amounts of data on a real-time basis and use big data software to perform analysis in real time. Imagine the volume of data collected and analyzed by Amazon, Walmart, banks, financial institutions, and online order services. The massive data collected is used for analysis and making quick decisions. Businesses are now run using data, and major business decisions are data-driven. The volume, variety, and speed with which data are now collected by different means are posing challenges in storage, data preparation, data cleaning, and analysis. To meet these challenging demands, different methods and software are being devised. The emergence of data science, data analytics, and business intelligence is the response to these challenges and needs. The techniques of machine learning and artificial intelligence that require large volumes of data are now becoming a necessity in this age of data. Data Science and the different areas of data science are now among the most sought-after skills in industries. We will explore several tools and technologies of data science in this book.

Some Other Sources of Data

- **Government agencies:** Government agencies, such as the Bureau of Labor Statistics, collect data on travel, health care, economic measures, unemployment, interest rates, etc.
- **Experimental design:** An experiment is any process that generates data. Design of experiment involves changing the values of input variables and observing their effect on the output or response variable of interest. The process provides useful data that can be used to optimize or troubleshoot a system.
- **Telephone /mail surveys:** Data can be obtained using telephone or mail surveys. Telephone surveys are inexpensive and very commonly used data collection method. The drawback is that it may generate low response rate as people may refuse to take part in the survey. The data collection method using mail questionnaire usually has the lowest response rate.
- **Processes:** Data are also collected from processes including manufacturing and service systems and processes. Data may directly be obtained from the process or experiments may be designed to obtain the needed information.

Analyzing Data Using Different Tools

The data, when collected, are usually referred to as raw data. Raw data are not analyzed. There is always a goal or objective behind collecting data. The raw data must be processed.

Data Related Terms Applied to Analytics

Big Data: Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications [Wikipedia].

As per **O'Reilly media**:

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it. **O'Reilly Media** made Big Data popular.

Gartner, who was credited with the 3 'V's of Big Data, classified big data as:

High-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation.

Gartner is referring to the size of data (large volume), the speed with which the data is being generated (velocity), and the different types of data (variety), and this seemed to align with the combined definition of Wikipedia and O'Reilly Media.

According to **Mike Gualtieri of Forrester**, the 3 'V's mentioned by Gartner are just measures of data. Mike insisted that Forrester's definition is more actionable. And that definition is:

Big Data is the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers.

Data mining: Data mining is about finding meaningful patterns and deriving insights in large sets of data using sophisticated pattern recognition techniques. It is closely related to Analytics that we discussed earlier. In data mining, you mine the data to get analytics. To derive meaningful patterns, data miners use statistics and statistical modeling, machine learning algorithms, and artificial intelligence.

Data warehouse: A **data warehouse (DW or DWH)**, or **enterprise data warehouse (EDW)**, is a system for storing, reporting, and analysis of huge amounts of data. The purpose of a DW is to create reports and perform analytics, which are core components of **Business Intelligence**. DWs are central repositories used to store and integrate current and historical data from one or many sources. The data are readily available and are used for creating analytical and visual reports throughout the enterprise. The data stored in the warehouse may be used for creating reports and performing analytics for the different operations in an enterprise, including sales, finance, marketing, engineering, and others. Before performing analyses on the data, cleansing, transformation, and data quality are critical issues.

Structured vs. Unstructured Data: Structured vs. Unstructured Data are the ‘Volume’ and ‘Variety’ – the ‘V’s of Big Data. Structured data is data that can be stored in relational databases. This type of data can be analyzed and organized in such a way that it can be related to other data via tables. Unstructured data cannot be directly put in databases or analyzed or organized directly. Some examples are email/text messages, social media posts, and recorded human speech, etc.

Data quality

Data quality is affected by the way data is collected, entered in the system, stored and managed. Efficient and accurate storage (data warehouse), cleansing, and data transformation are critical for assuring data quality. The process of verifying the reliability and effectiveness of data is sometimes referred to as Data quality assurance (DQA). The effectiveness, reliability, and success of business analytics (BA) and business intelligence (BI) depend on acceptable data quality.

The following are important considerations in assuring data quality. Aspects of data quality include: [<http://searchdatamanagement.techtarget.com/definition/data-quality>]

- Accuracy
- Completeness
- Update status
- Relevance
- Consistency across data sources
- Reliability
- Appropriate presentation
- Accessibility

Within an organization, acceptable data quality is crucial to operational and transactional processes.

Summary

Data science is about the study of data. Data are of various types and are collected using different means. Data are not only numbers but are classified as text data or qualitative or categorical data and also numerical or quantitative data. This chapter explained the types of data and their classification with examples. Data can also be collected through audio and video devices. Companies collect massive amounts of data. The volume of data collected and analyzed by businesses is so large that it is referred to as “Big Data.” The volume, variety, and the speed (velocity) with which data are collected require specialized tools and techniques including specially designed big data software for analysis. In many cases, the data must be analyzed in real time to make effective business decisions. Data Science and associated analytical tools are being designed to meet these challenges.