Unit-II

Faculty: D Sai Kumar

Dept. of CSE, UCE, OU

Techniques : Introduction

- There are many tools and techniques that a data scientist is expected to know or acquire as problems arise.

- It is hard to separate tools and techniques.

Data Analysis and Data Analytics

Data Analysis:

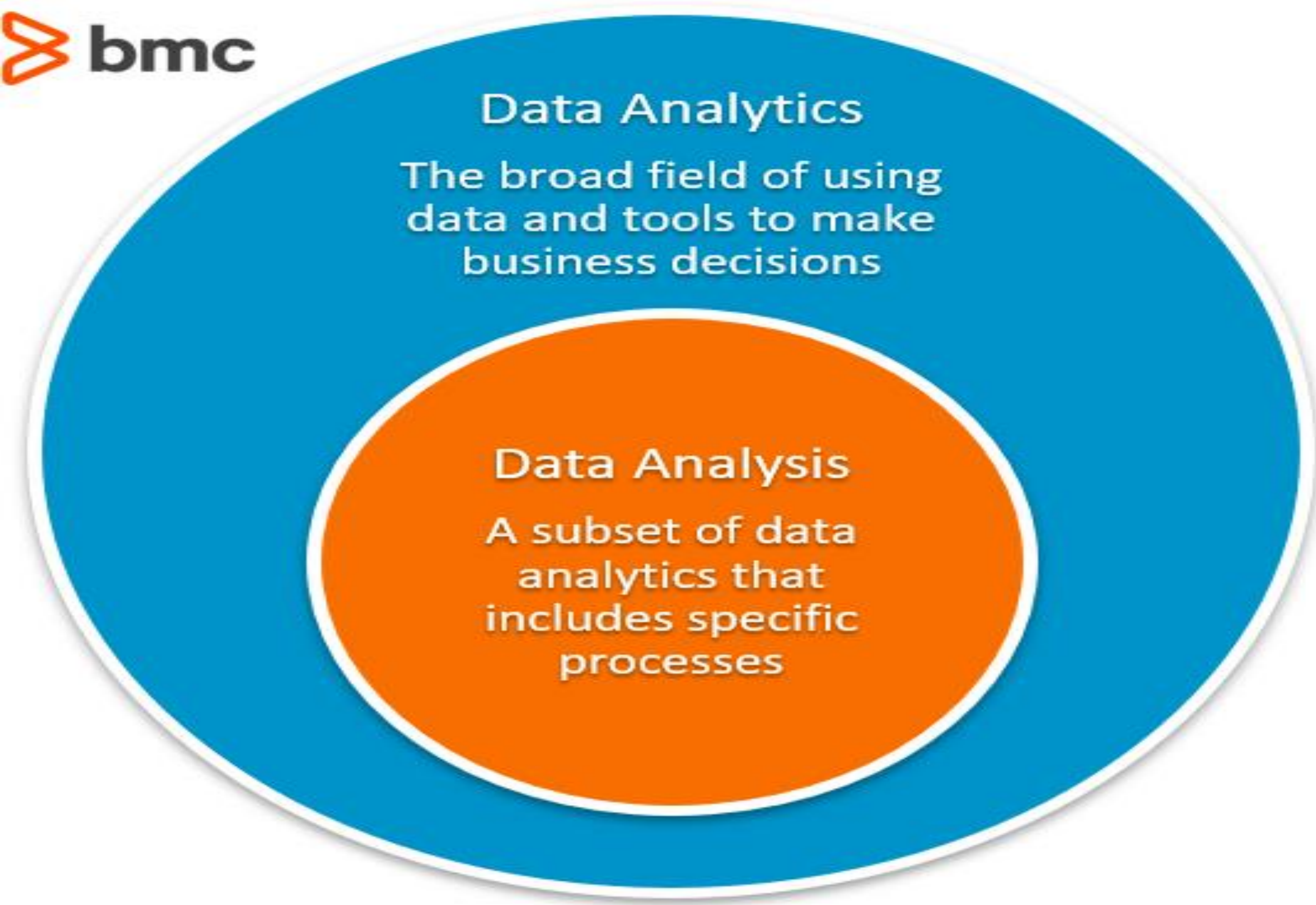
- Data analysis is the process of examining raw data to draw conclusions.
- It involves cleaning, transforming, and summarizing data to identify patterns, trends, and anomalies.
- Data analysis is often used to understand past events and current situations.

Example: summarizing sales trends.

Data Analytics:

- Data analytics is a broader field that includes data analysis and goes further to make predictions and inform decisions.
- It leverages advanced techniques like machine learning and artificial intelligence to anticipate future outcomes and optimize business processes.

Example: A retail store analyzes sales data (data analysis) and uses it to forecast demand (data analytics).



Data Analytics

The broad field of using data and tools to make business decisions

Data Analysis

A subset of data analytics that includes specific processes

- One way to understand the difference between analysis and analytics is to think in terms of past and future.
- Analysis looks backwards, providing marketers with a historical view of what has happened.
- Analytics, on the other hand, models the future or predicts a result.
- Analytics makes extensive use of mathematics and statistics and the use of descriptive techniques and predictive models to gain valuable knowledge from data to recommend action or to guide decision-making in a business context.

Descriptive Analysis

- Descriptive analysis focuses on understanding **"what is happening now"** based on incoming data.
- It quantitatively summarizes key features of a dataset, serving as the first step in data analysis, especially for large datasets like census data.
- The key points about descriptive analysis:
    - ➢ Typically, it is the first kind of data analysis performed on a dataset.
    - ➢ Usually it is applied to large volumes of data, such as census data.
    - ➢ Description and interpretation processes are different steps.
- Descriptive analysis facilitate analyzing and summarizing the data and are thus instrumental to processes inherent in data science.

Variables

- Before Data is processed or analyzed, we first need to **capture** it and **represent** it.
- Variables specifically does this for us, they capture and represent data that can be further processed or Analyzed. A variable is a label we give to our data.
- A variable is a characteristic or attribute that can be measured or observed. It's something that can vary or take on different values.

For example :

Name	Age	Student
Gautam	21	Yes
Ramesh	44	No
Sujit	29	Yes

Here in the above table a persons age is Numeric variable, which represents that person, also Student is as well a variable but a categorical type.

Thus Numerical variables which are more frequently used in Data science can be further categorized as being

- Counting
- Ranking
- Scaling

etc.. depending how they will be used for analysis

Types:

- **Qualitative (Categorical):** Describe qualities or categories.
    - **Nominal:** Unordered categories (e.g., colors, gender, types of fruit). No inherent ranking.
    - **Ordinal:** Ordered categories (e.g., education level (high school, bachelor's, master's), satisfaction rating (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied)). There's a ranking, but the intervals between categories may not be equal.

- **Quantitative (Numerical):** Represent numerical values.
    - **Discrete:** Countable values (e.g., number of children, number of cars). Usually whole numbers.
    - **Continuous:** Measurable values (e.g., height, weight, temperature). Can take on any value within a range.

- All of these categories of variables are fine when we are dealing with one variable at a time and doing descriptive analysis. But when we are trying to connect multiple variables or using one set of variables to make predictions about another set, we may want to classify them with some other names.

- A variable that is thought to be controlled or not affected by other variables is called an **independent variable**.
- A variable that depends on other variables (most often other independent variables) is called a **dependent variable**.
- In the case of a prediction problem, an independent variable is also called a **predictor variable** and a dependent variable is called an **outcome variable**.

E.g: Independent variable and Dependent variable

Patient ID	Tumor Size (mm)	Cancer
1	15	Yes
2	8	No
3	22	Yes
4	5	No
5	18	Yes
6	10	No
7	25	Yes
8	7	No
9	12	Yes
10	3	No

Independent Variable: Tumor Size (mm) - This variable is believed to potentially influence the presence or absence of cancer.

Dependent Variable: Cancer - This is the outcome we are trying to predict or understand based on the tumor size.

Frequency Distribution

- Once data is collected , that need to be visualized (presented) for describing it.
- How many times a data score (data items) are occurring in the data can be visualized (presented) best by plotting graphs (different types)
- Data items occurring is referred to by its frequency in data, that is “Data Frequency”.
- Frequency Distribution explains, how this frequent data items are spread across.
- A frequency distribution shows how often each value (or range of values) of a variable occurs in a dataset. It can be presented as a table or a graph (histogram, bar chart, etc.).

- **Histogram** : Histograms plot values of observations on the horizontal axis, with a bar showing how many times each value occurred in the dataset.

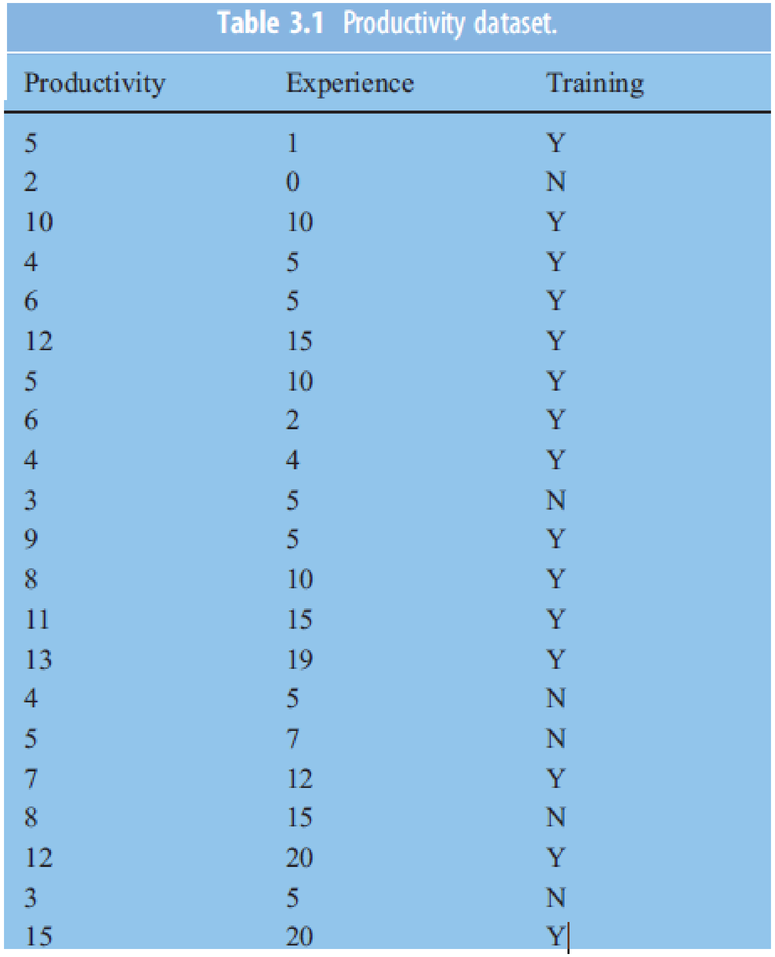
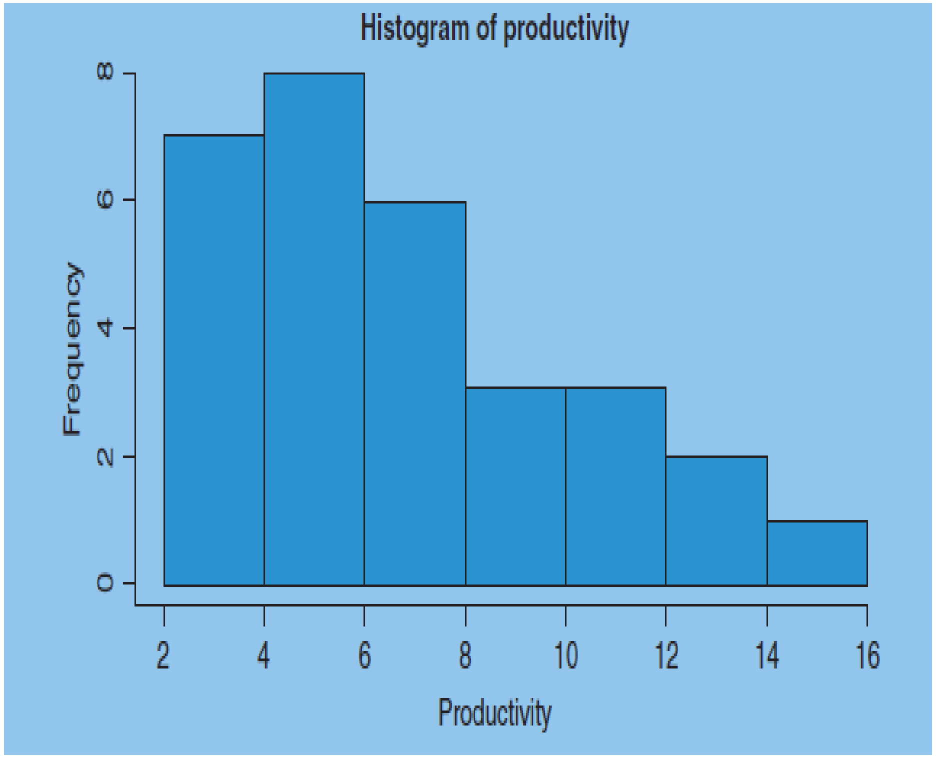
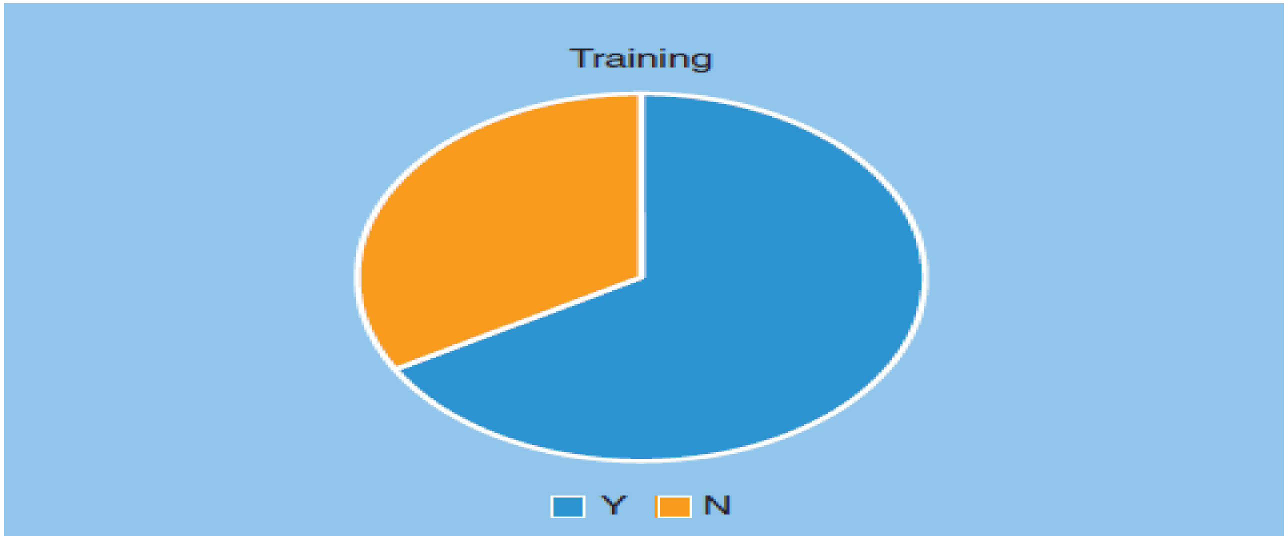


Table 3.1 Productivity dataset.

Productivity	Experience	Training
5	1	Y
2	0	N
10	10	Y
4	5	Y
6	5	Y
12	15	Y
5	10	Y
6	2	Y
4	4	Y
3	5	N
9	5	Y
8	10	Y
11	15	Y
13	19	Y
4	5	N
5	7	N
7	12	Y
8	15	N
12	20	Y
3	5	N
15	20	Y

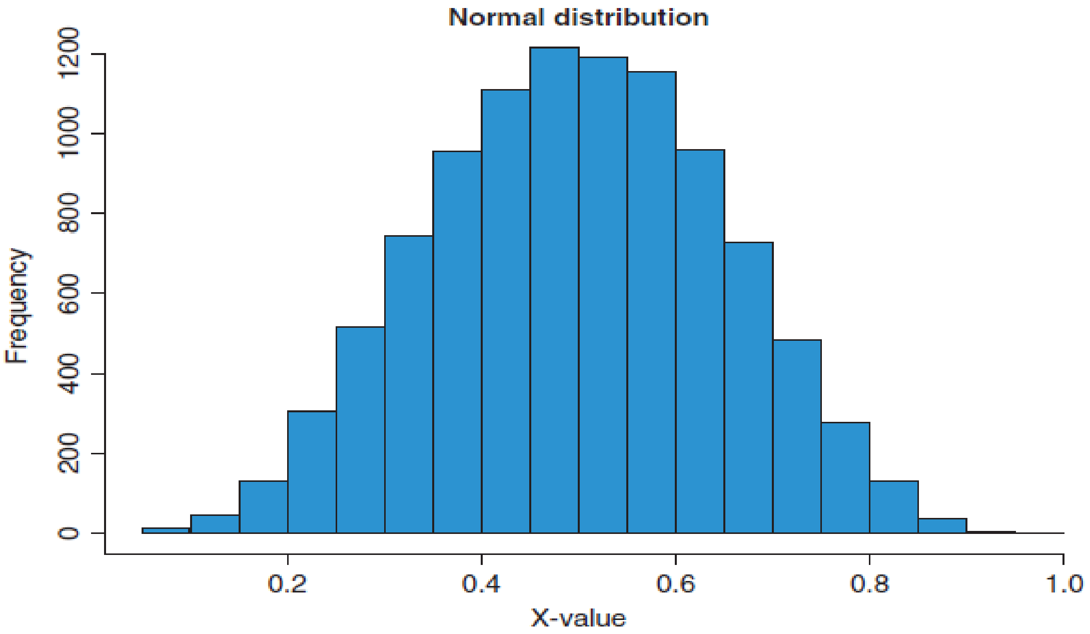


- Histogram worked fine for numerical data, but what about categorical data? In other words, how do we visualize the data when it’s distributed in a few finite categories? We have such data in the third column called “Training.” For that, we can create a pie chart



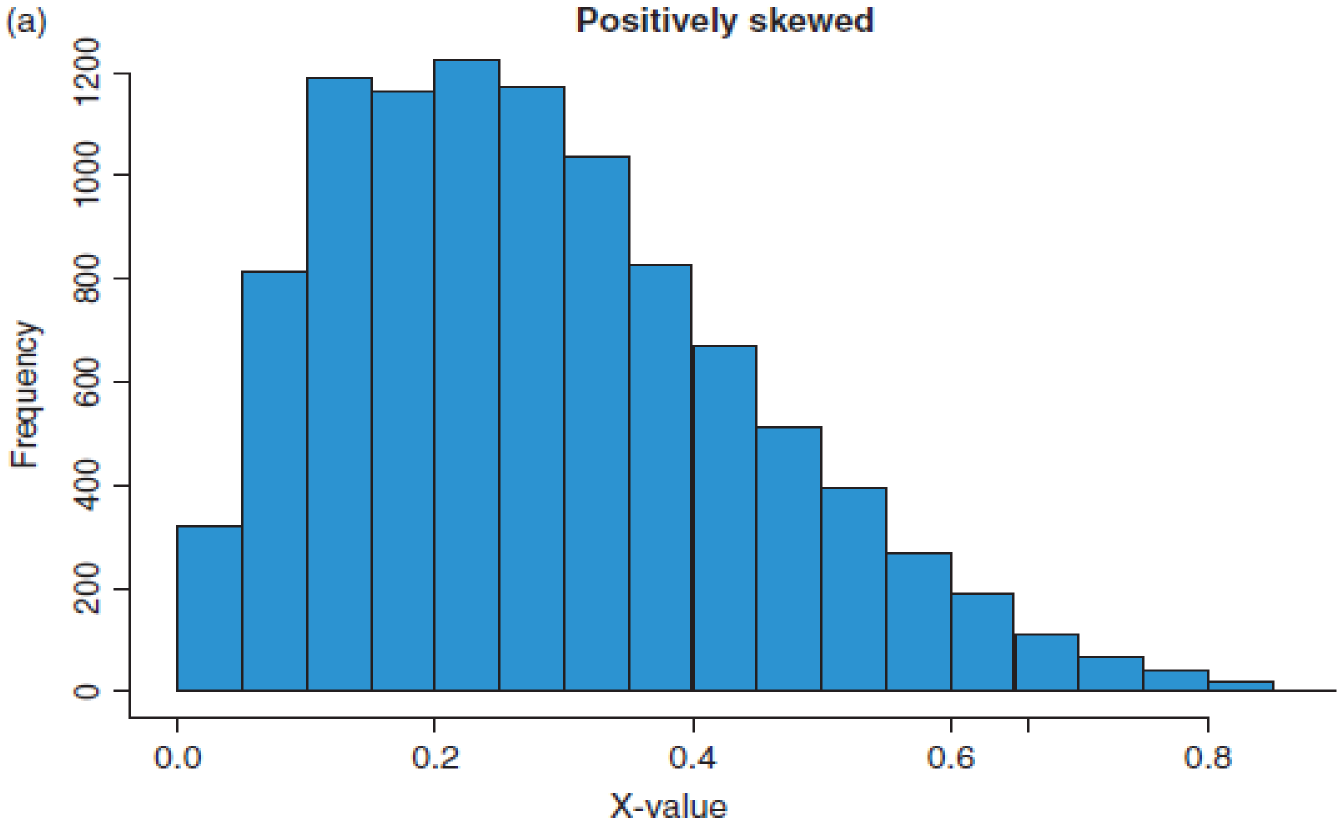
Pie chart showing the distribution of “Training” in the Productivity data.

- We will often be working with data that are numerical and we will need to understand how those numbers are spread. For that, we can look at the nature of that distribution.
- It turns out that, if the data is normally distributed, various forms of analyses become easy and straightforward.
- **Normal Distribution**. In an ideal world, data would be distributed symmetrically around the center of all scores. Thus, if we drew a vertical line through the center of a distribution, both sides should look the same. This so-called normal distribution, is characterized by a bell-shaped curve, an example of which is shown in Figure

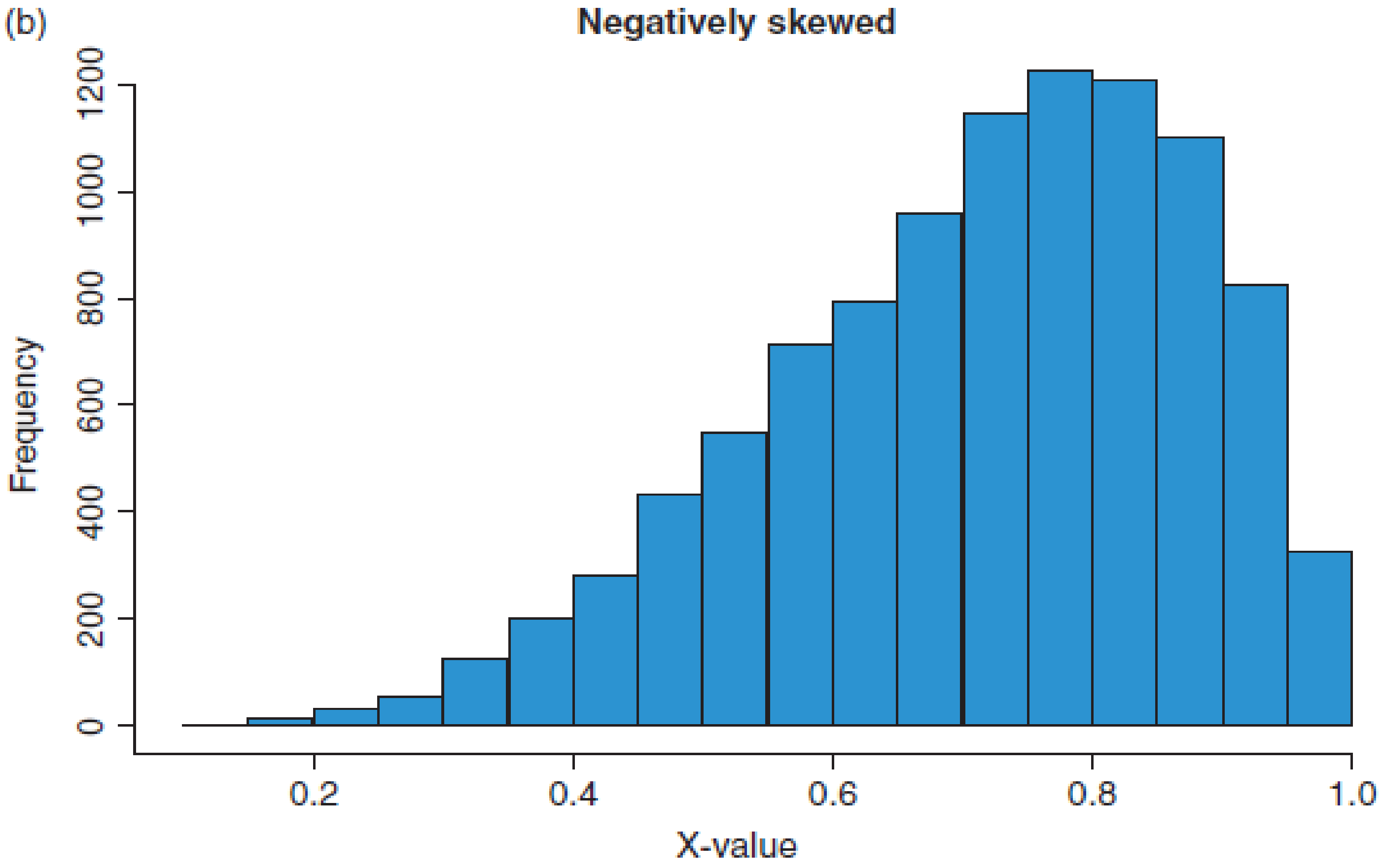


- There are two ways in which a distribution can deviate from normal:
    - Lack of symmetry (called skew)
    - Pointiness (called kurtosis)

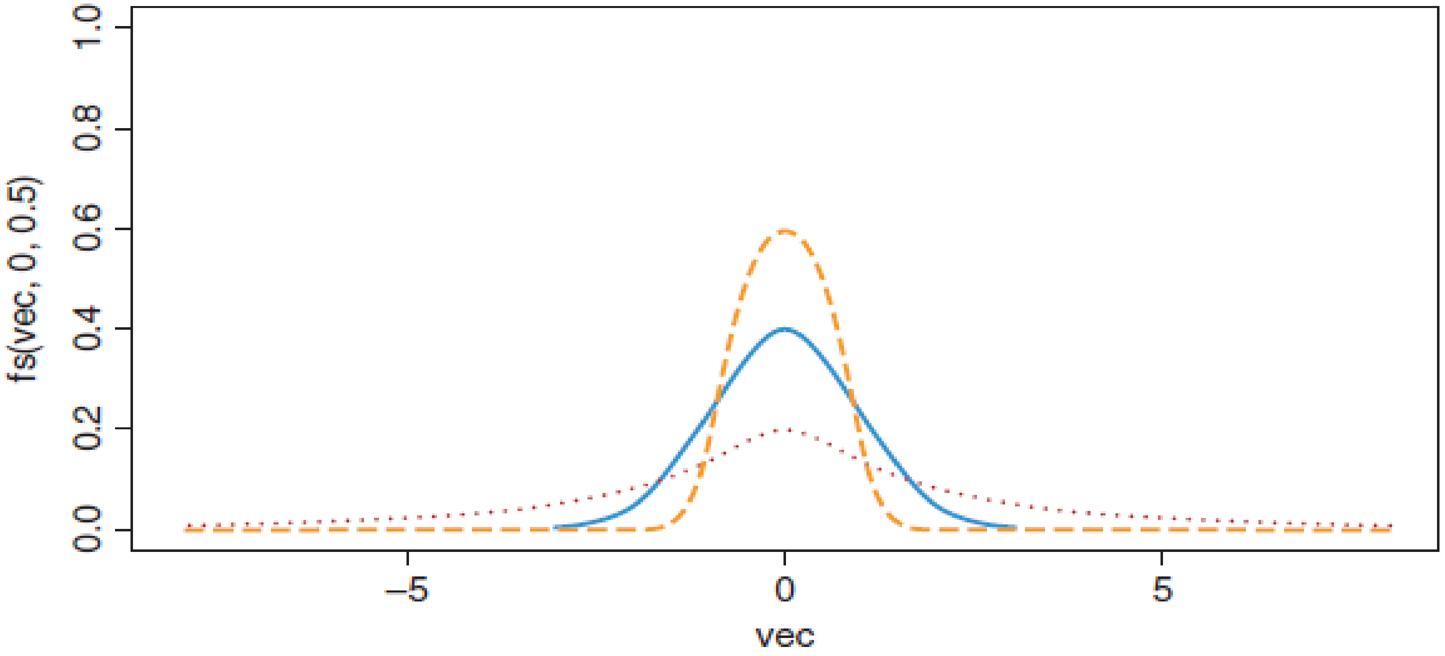
Positive Skew (Right Skew):



Negative Skew (Left Skew):



- **Kurtosis**, on the other hand, refers to the degree to which scores cluster at the end of a distribution (platykurtic) and how “pointy” a distribution is (leptokurtic), as shown in Figure



Examples of different kurtosis in a distribution (orange dashed line represents leptokurtic, blue solid line represents the normal distribution, and red dotted line represents platykurtic).

Measures of Centrality

- Often, one number can tell us enough about a distribution. This is typically a number that points to the “center” of a distribution. In other words, we can calculate where the “center” of a frequency distribution lies, which is also known as the **central tendency**.

- There are three measures commonly used:

1. Mean
2. median, and
3. mode.

Mean

- You have come across this before even if you have never done statistics.
- Mean is commonly known as average, though they are not exactly synonyms.
- Mean is most often used to measure the central tendency of continuous data as well as a discrete dataset.
- If there are n number of values in a dataset and the values are x_1, x_2, . . ., x_n, then the mean is calculated as

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}.$$

- There is a significant drawback to using the mean as a central statistic: it is susceptible to the influence of outliers(Extreme Values).
- Also, mean is only meaningful if the data is normally distributed, or at least close to looking like a normal distribution.

Example: Ages: 25, 30, 35, 40, 45.

Mean = (25+30+35+40+45) / 5 = 35

Median

- The median is the middle score for a dataset that has been sorted according to the values of the data.
- With an even number of values, the median is calculated as the average of the middle two data points.
- **Example:** Ages: 25, 30, 35, 40, 45.

  Median = 35.

  Ages: 25, 30, 35, 40.

  Median = (30+35)/2 = 32.5

Robustness: Less sensitive to outliers than the mean

Mode

- The mode is the most frequently occurring value in a dataset.
- On a histogram representation, the highest bar denotes the mode of the data.
- Normally, mode is used for categorical data; for example, for the Training component in the Productivity dataset, the most common category is the desired output.

Example: Ages: 25, 30, 30, 35, 40.

Mode = 30.

Usefulness: Useful for categorical data and multimodal distributions (distributions with multiple peaks). A dataset can have no mode, one mode (unimodal), or multiple modes (bimodal, trimodal, etc.).

Dispersion of a Distribution

- Always measures of centrality will not give sufficient idea of the distribution, in that we need to look at the spread of the distribution.

Dispersion can be measured using

- Range
- Interquartile Range
- Variance
- Standard Deviation

- **Range** :The easiest way to look at the dispersion is to take the largest score and subtract it from the smallest score. This is known as the range

  Example: Ages: 25, 30, 35, 40, 45.

  Range = 45 - 25 = 20.

- **Limitation:** Only considers the extremes, not the distribution of the data in between.

Interquartile Range

- One way around the range’s disadvantage is to calculate it after removing extreme values. One convention is to cut off the top and bottom one-quarter of the data and calculate the range of the remaining middle 50% of the scores.

This is known as the interquartile range.

- Order the data. Q1 is the median of the lower half. Q3 is the median of the upper half. IQR = Q3 - Q1.

Example: Ages: 25, 30, 35, 40, 45.

Q1 = 30, Q3 = 40.

IQR = 40 - 30 = 10.

- Less sensitive to outliers than the range.

Variance

- The variance is a measure used to indicate how spread out the data points are.
- To measure the variance, the common method is to pick a center of the distribution, typically the mean, then measure how far each data point is from the center.
- If the individual observations vary greatly from the group mean, the variance is big; and vice versa.
- Here, it is important to distinguish between the variance of a population and the variance of a sample.
- They have different notations, and they are computed differently.
- The variance of a population is denoted by σ^2; and
- the variance of a sample by s^2.

- The variance of a population is defined by the following formula:

$$\sigma^2 = \frac{\sum (X_i - X)^2}{N},$$

where σ^2 is the population variance, X is the population mean, Xi is the ith element from the population, and N is the number of elements in the population

- The variance of a sample is defined by a slightly different formula:

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

- where s^2 is the sample variance, x is the sample mean, xi is the ith element from the sample, and n is the number of elements in the sample.
- Using this formula, the variance of the sample is an unbiased estimate of the variance of the population.

Standard Deviation

- There is one issue with the variance as a measure. It gives us the measure of spread in units squared.
- So, for example, if we measure the variance of age (measured in years) of all the students in a class, the measure we will get will be in year^2
- However, practically, it would make more sense if we got the measure in years (not years squared).
- For this reason, we often take the square root of the variance, which ensures the measure of average spread is in the same units as the original measure. This measure is known as the **standard deviation**
- The formula to compute the standard deviation of a sample is

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

Diagnostic Analytics

- Diagnostic analytics delves deeper than descriptive analysis to understand *why* something happened. It focuses on identifying the root causes of events or trends.
- Sometimes this type of analytics when done hands-on with a small dataset is also known as **causal analysis**, since it involves at least one cause (usually more than one) and one effect.

- There are various types of techniques available for diagnostic or causal analytics. Among them, one of the most frequently used is **correlation**.

Benefits of Diagnostic Analytics:

- Improved Decision-Making
- Enhanced Problem-Solving
- Resource Optimization
- Competitive Advantage

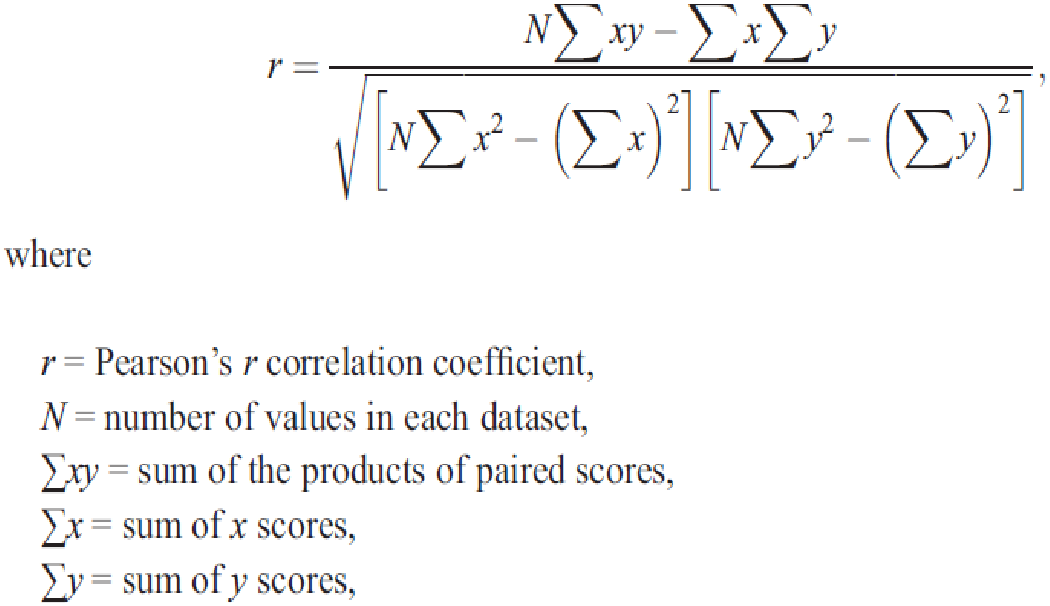
Correlations

- Correlation is a statistical analysis that is used to measure and describe the strength and direction of the relationship between two variables.
  - Strength indicates how closely two variables are related to each other,
  - direction indicates how one variable would change its value as the value of the other variable changes.

- Correlation is a simple statistical measure that examines how two variables change together over time.
- for example, “umbrella” and “rain.”
- If someone who grew up in a place where it never rained saw rain for the first time, this person would observe that, whenever it rains, people use umbrellas. They may also notice that, on dry days, folks do not carry umbrellas.

- By definition, “rain” and “umbrella” are said to be correlated!

- An important statistic, the **Pearson's r correlation**, is widely used to measure the degree of the relationship between linear related variables.
- The following formula is used to calculate the Pearson's r correlation:



$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}},$$

where

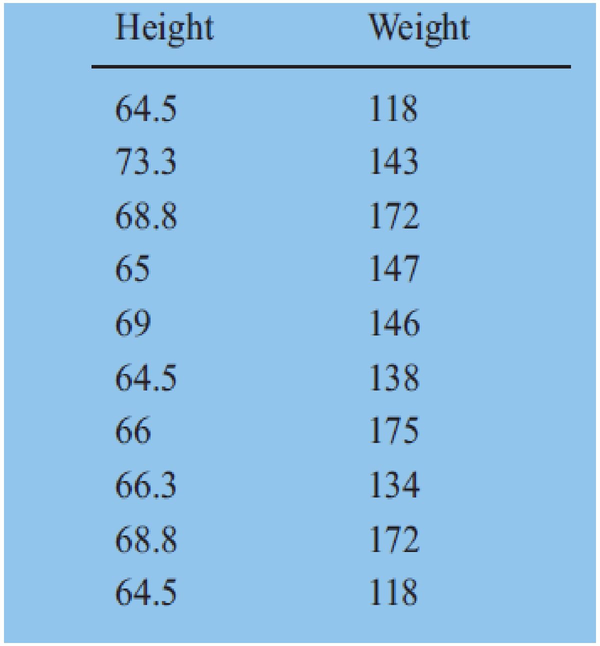
r = Pearson's r correlation coefficient,
N = number of values in each dataset,
$\sum xy$ = sum of the products of paired scores,
$\sum x$ = sum of x scores,
$\sum y$ = sum of y scores,

Σx2 = sum of squared x scores, and
Σy2 = sum of squared y scores.6

The value of correlation lies between -1 and +1.

Example:

- Let us use the formula and calculate Pearson's r correlation coefficient for the height– weight pair with the data provided.



Height	Weight
64.5	118
73.3	143
68.8	172
65	147
69	146
64.5	138
66	175
66.3	134
68.8	172
64.5	118

N = number of values in each dataset = 10

Σ xy = sum of the products of paired scores = 98,335.30

Σ x = sum of x scores = 670.70

Σ y = sum of y scores = 1463

Σ x^2 = sum of squared x scores = 45,058.21

Σ y^2 = sum of squared y scores = 218,015

The Pearson's r correlation formula gives us 0.39 (approximated to two decimal places) as the correlation coefficient.

This indicates two things: (1) "height" and "weight" are positively related, which means that, as one goes up, so does the other; and (2) the strength of their relation is medium.

Predictive Analytics

- Uses historical data to predict *future* outcomes. Employs statistical modeling and machine learning.
- These analytics are about understanding the future using the data and the trends we have seen in the past, as well as emerging new contexts and processes
- An example is trying to predict how people will spend their tax refunds based on how consumers normally behave around a given time of the year (past data and trends), and how a new tax policy (new context) may affect people's refunds.

Predictive analytics is done in stages.

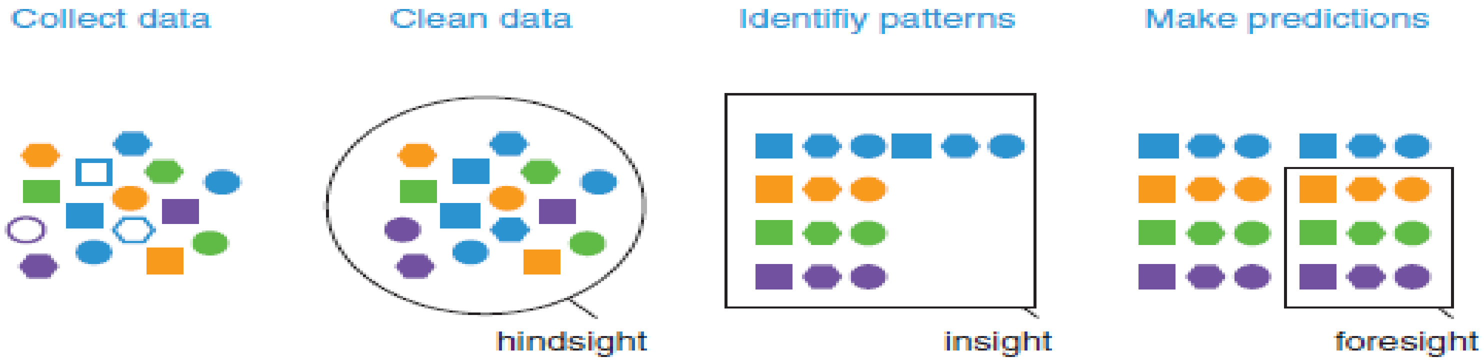
1. First, once the data collection is complete, it needs to go through the process of cleaning

2. Cleaned data can help us obtain hindsight in relationships between different variables. Plotting the data (e.g., on a scatterplot) is a good place to look for hindsight.

3. Next, we need to confirm the existence of such relationships in the data. This is where regression comes into play. From the regression equation, we can confirm the pattern of

distribution inside the data. In other words, we obtain insight from hindsight.

4. Finally, based on the identified patterns, or insight, we can predict the future, i.e.,



Prescriptive Analytics

- Recommends actions to take to achieve desired outcomes. Builds upon predictive analytics.
- Prescriptive analytics is the area of business analytics dedicated to finding the best course of action for a given situation.
- A process-intensive task, the prescriptive approach analyzes potential decisions, the interactions between decisions, the influences that bear upon these decisions, and the bearing all of this has on an outcome to ultimately prescribe an optimal course of action in real time.
- Prescriptive analytics can also suggest options for taking advantage of a future opportunity or mitigate a future risk and illustrate the implications of each.
- In practice, prescriptive analytics can continually and automatically process new data to improve the accuracy of predictions and provide advantageous decision options.

- Specific techniques used in prescriptive analytics include optimization, simulation, game theory and decision-analysis methods.
- Prescriptive analytics can be really valuable in deriving insights from given data, but it is largely not used.
- According to Gartner, 13% of organizations are using predictive analytics, but only 3% are using prescriptive analytics.

- There are two more categories of data analysis techniques that are different from the above-mentioned four categories

➢ Exploratory analysis and

➢ Mechanistic analysis.

Exploratory Analysis

- Exploratory Data Analysis (EDA) is an approach used to **uncover previously unknown relationships** within a dataset. It relies heavily on **data visualization techniques** to help researchers identify patterns, trends, and anomalies in data.
- In other words, we are asked to provide an answer without knowing the question! This is where we go for an exploration.
- When we lack a clear question or a hypothesis, plotting the data in different forms could provide us with some clues regarding what we may find or want to find in the data.
- Such insights can then be useful for defining future studies/questions, leading to other forms of analysis.
- Thus, exploratory analysis is not a mere collection of techniques; rather, it offers a philosophy as to how to dissect a dataset

    what to look for;

    how to look;

    and how to interpret the outcomes.

Example can be census data which will have dozens of variables.

Suppose you are looking for a state with highest population then you will go for **“Descriptive Analysis”**

If you are predicting something like a state with less number of migrants then you may go for “**Prescriptive or Predictive Analysis”.**

But if you want to know interesting insights from data then you have to consider

- huge data
- lot of variables

Then you end up trying “**Exploratory Analysis”**

Mechanistic Analysis

- Mechanistic analysis involves understanding the exact changes in variables that lead to changes in other variables for individual objects.
- Focuses on understanding the *cause-and-effect* relationships between variables.
- For instance, we may want to know how the number of free doughnuts per employee per day affects employee productivity. Perhaps by giving them one extra doughnut we gain a 5% productivity boost, but two extra doughnuts could end up making them lazy (and diabetic)!

- So need a optimum level where focus can be made.

Regression

- Mechanistic Analysis is studying a relationship between two variables. Such relationships are often explored using **regression**.
- **Regression** analysis is a process for estimating the relationships among variables.
- A statistical method for modeling the relationship between a dependent variable and one or more independent variables.

Example: Predicting house prices based on size, location, and age

Regression vs Correlation

- Correlation by itself does not provide any indication of how one variable can be predicted from another.
- Regression provides this crucial information.

- Linear regression, the most common form of regression used in data analysis, assumes this relationship to be linear. In other words, the relationship of the predictor variable(s) and outcome variable can be expressed by a straight line.
- If the predictor variable is represented by x, and the outcome variable is represented by y, then the relationship can be expressed by the equation

$$y = \beta_0 + \beta_1 x,$$

- where β_1 represents the slope of the x, and β_0 is the intercept or error term for the equation.
- What linear regression does is estimate the values of β_0 and β_1 from a set of observed data points, where the values of x, and associated values of y, are provided. So, when a new or previously unobserved data point comes where the value of y is unknown, it can fit the values of x, β_0, and β_1 into the above equation to predict the value of y.

- From statistical analysis, it has been shown that the slope of the regression β_1 can be expressed by the following equation:

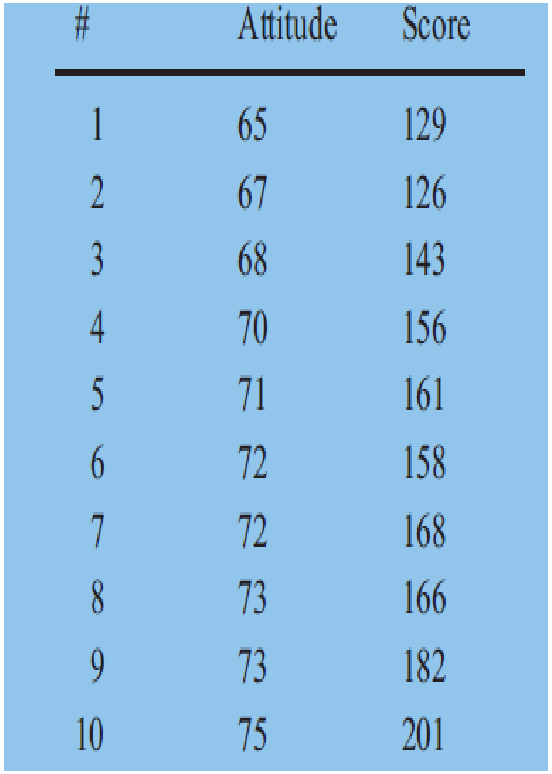
$$\beta_1 = r\frac{\mathrm{sd}_y}{\mathrm{sd}_x},$$

- where r is the Pearson's correlation coefficient, and sd represents the standard deviation of the respective variable as calculated from the observed set of data points. Next, the value of the error term can be calculated from the following formula:

$$\beta_0 = \bar{y} - \beta_1\bar{x},$$

- where $\bar{y}$ and $\bar{x}$ represent the means of the y and x variables, respectively.

- We use the attitude dataset, The first variable, attitude, represents the amount of positive attitude of the students who have taken an examination, and the score represents the marks scored by the participants in the examination.



#	Attitude	Score
1	65	129
2	67	126
3	68	143
4	70	156
5	71	161
6	72	158
7	72	168
8	73	166
9	73	182
10	75	201

Here attitude is going to be the predictor variable, and what regression would be able to do is to estimate the value of score from attitude

From the data, Pearson's correlation coefficient r can be calculated as 0.94. The standard deviations of x (attitude) and y (score) are 3.10 and 22.80, respectively

β_1 = 0.94*(22.80/ 3.10) = 6.91

β_0 = 159 - (6:91 * 70:6) = - 328. 85

Now, say you have a new participant whose positive attitude before taking the examination is measured at 78. His score in the examination can be estimated at 210.13:

y = - 328.85 +(6.91 * 78) = 210.13

Statistics

- “Statistics” refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.
- The study of statistics involves math and relies upon calculations of numbers.
- But it also relies heavily on how the numbers are chosen and how data is interpreted.

Examples:

1. A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective.

2. 75% more interracial marriages are occurring this year than 25 years ago. Thus, our society accepts interracial marriages.

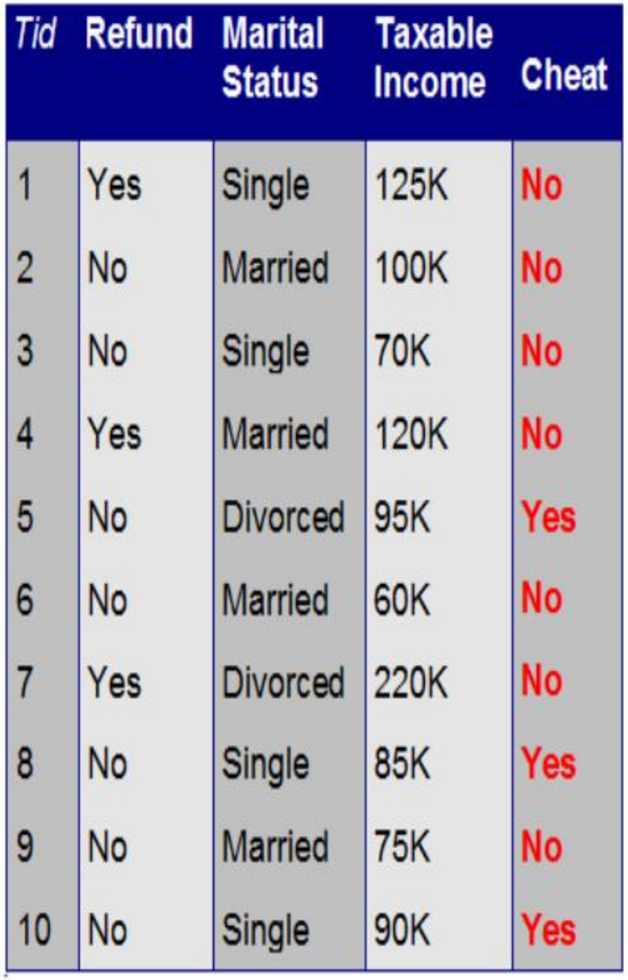
Understanding attributes and their types

- **Data sets** are made up of **data objects**.
- A **data object** represents an entity.
  - – Also called **sample, example, instance, data point, object, tuple**.
- Data objects are described by **attributes**.
- An **attribute** is a property or characteristic of a data object.
  - – Examples: eye color of a person, temperature, etc.
  - – Attribute is also known as **variable, field, characteristic, or feature**
- A collection of attributes describe an object.
- **Attribute values** are numbers or symbols assigned to an attribute.

A Data Object

Attributes

Objects



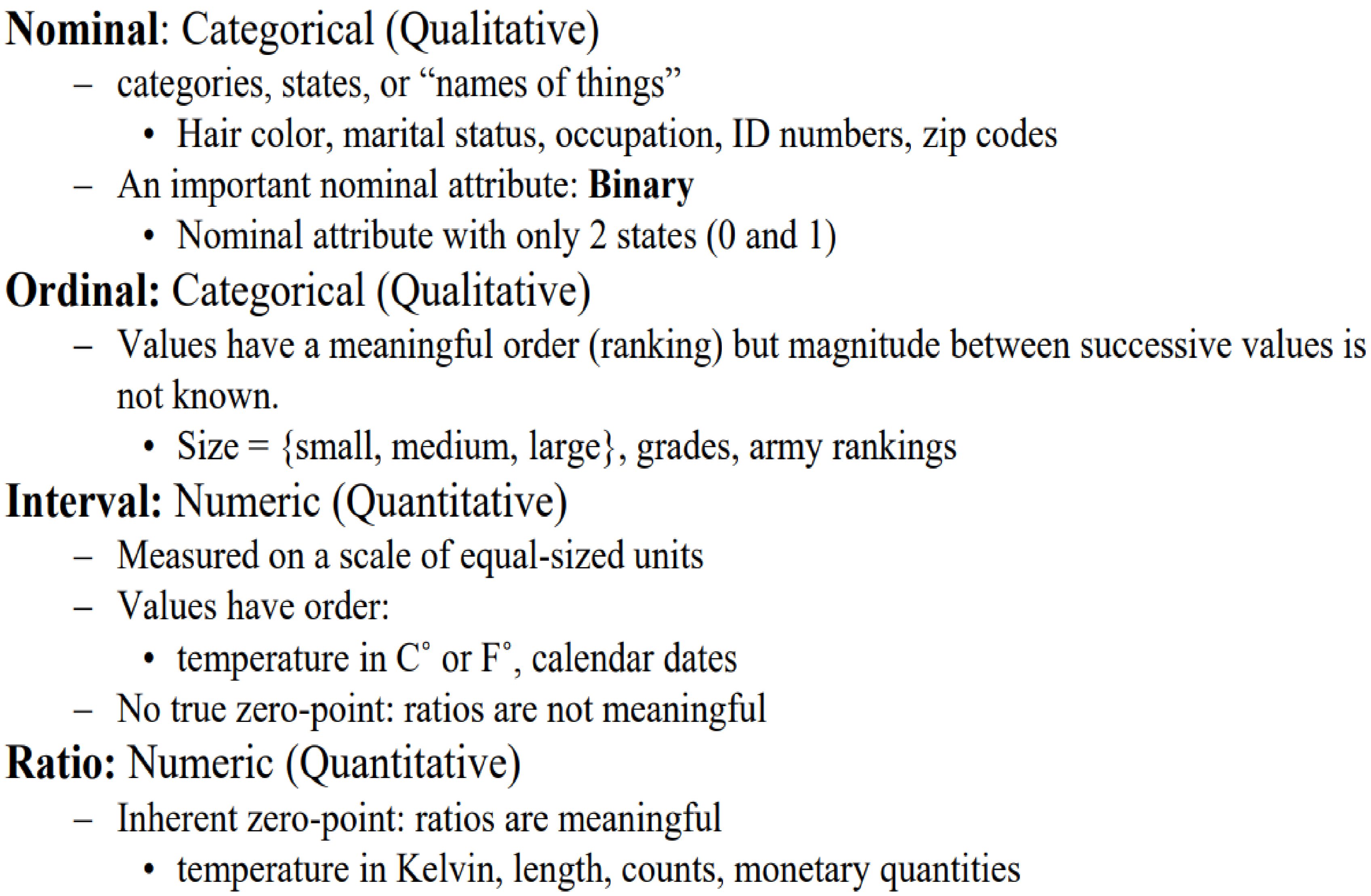
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- database rows ➔ data objects
- database columns ➔ attributes

Attributes

- **Attribute** (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
  - E.g., customer _ID, name, address
- **Attribute values** are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different; ID has no limit but age has a maximum and minimum value

Types of attributes



Nominal: Categorical (Qualitative)

- categories, states, or "names of things"
  - Hair color, marital status, occupation, ID numbers, zip codes
- An important nominal attribute: **Binary**
  - Nominal attribute with only 2 states (0 and 1)

Ordinal: Categorical (Qualitative)

- Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings

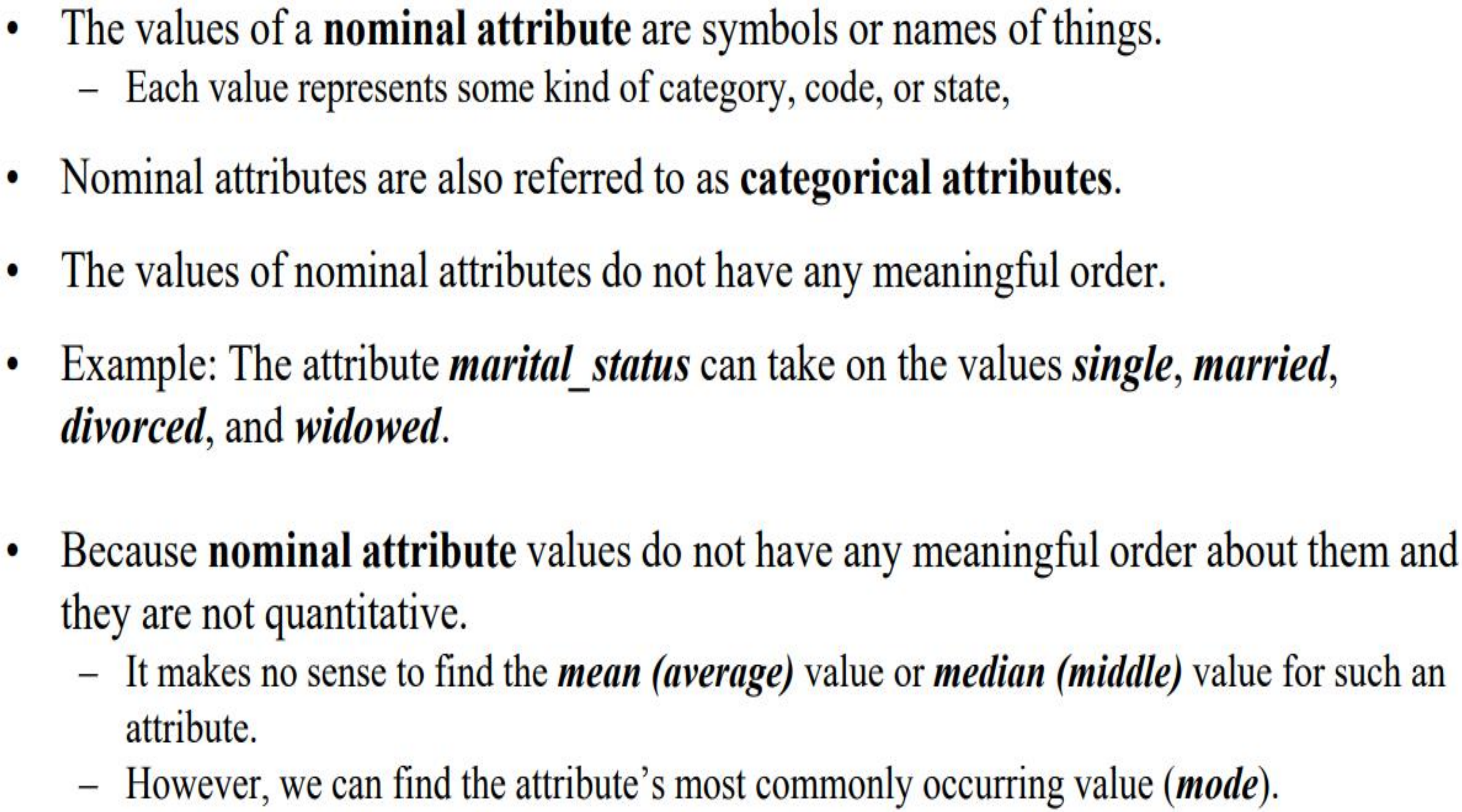
Interval: Numeric (Quantitative)

- Measured on a scale of equal-sized units
- Values have order:
  - temperature in C° or F°, calendar dates
- No true zero-point: ratios are not meaningful

Ratio: Numeric (Quantitative)

- Inherent zero-point: ratios are meaningful
  - temperature in Kelvin, length, counts, monetary quantities

Nominal Attributes



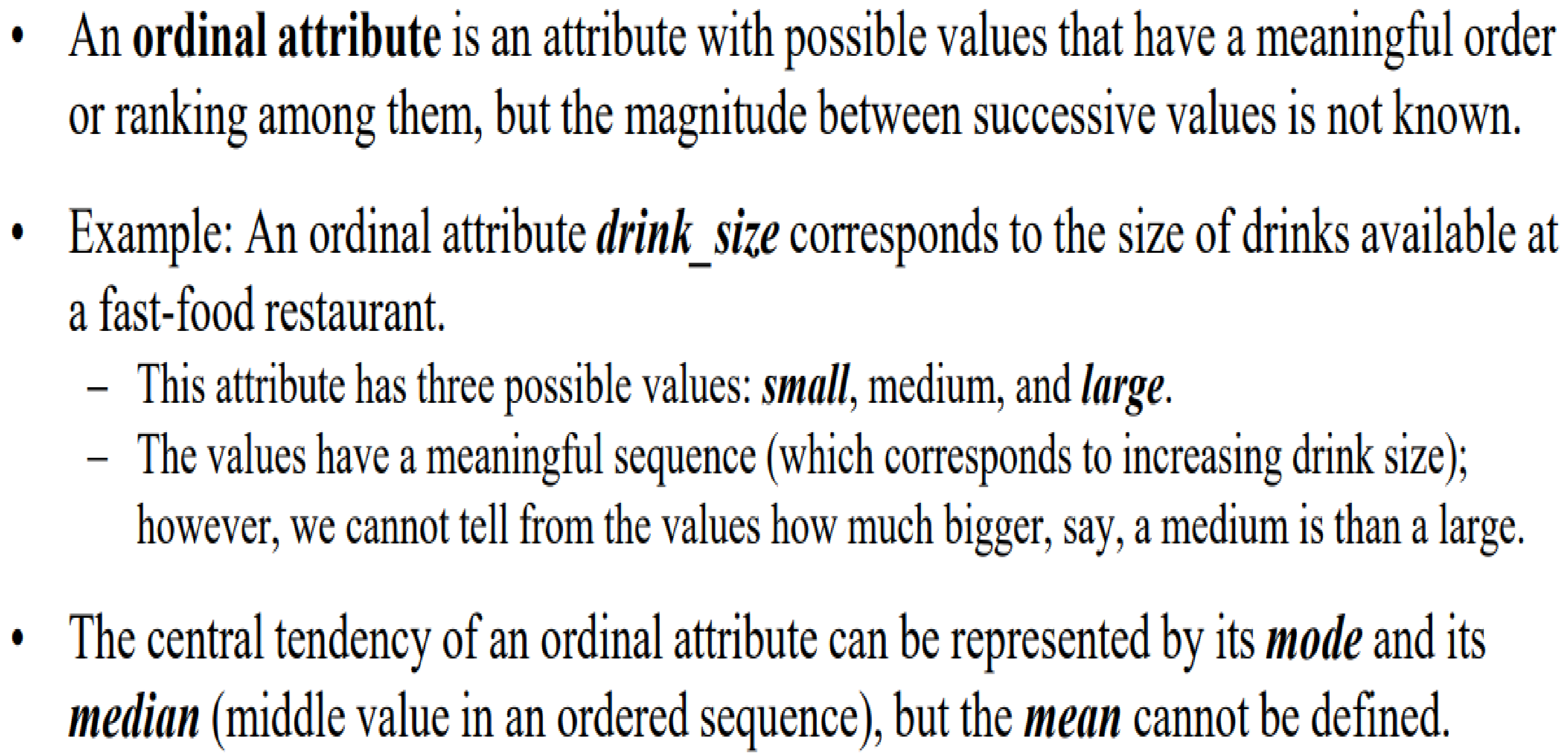
- The values of a **nominal attribute** are symbols or names of things.
  - Each value represents some kind of category, code, or state,
- Nominal attributes are also referred to as **categorical attributes**.
- The values of nominal attributes do not have any meaningful order.
- Example: The attribute ***marital_status*** can take on the values ***single***, ***married***, ***divorced***, and ***widowed***.
- Because **nominal attribute** values do not have any meaningful order about them and they are not quantitative.
  - It makes no sense to find the ***mean (average)*** value or ***median (middle)*** value for such an attribute.
  - However, we can find the attribute's most commonly occurring value (***mode***).

- A **binary attribute** is a special ***nominal attribute*** with only two states: 0 or 1.

- A binary attribute is **symmetric** if both of its states are equally valuable and carry the same weight.
  - Example: the attribute ***gender*** having the states ***male*** and ***female***.

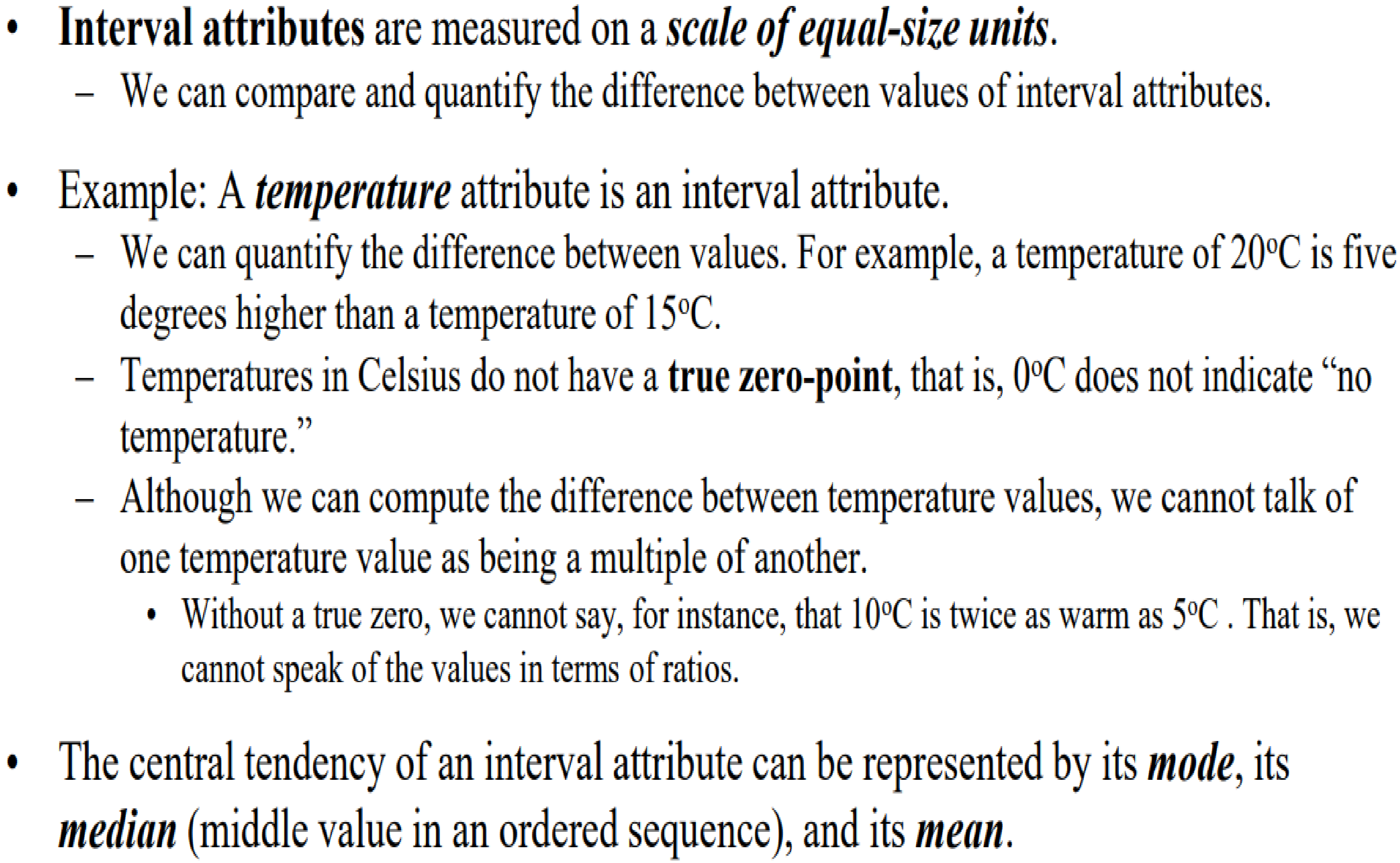
- A binary attribute is **asymmetric** if the outcomes of the states are not equally important.
  - Example: ***Positive*** and ***negative*** outcomes of a medical test for HIV.
  - By convention, we code the most important outcome, which is usually the rarest one, by 1 (e.g., HIV positive) and the other by 0 (e.g., HIV negative).

Ordinal Attributes



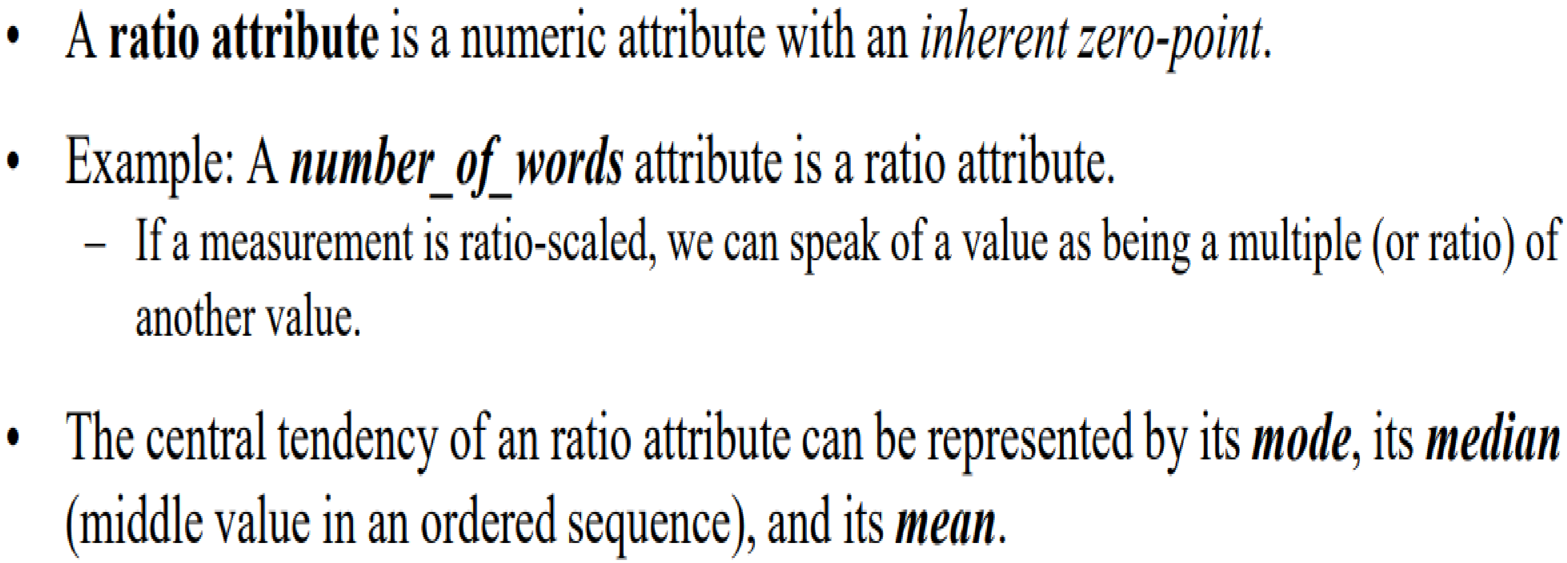
- An **ordinal attribute** is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.
- Example: An ordinal attribute ***drink_size*** corresponds to the size of drinks available at a fast-food restaurant.
  - This attribute has three possible values: ***small***, medium, and ***large***.
  - The values have a meaningful sequence (which corresponds to increasing drink size); however, we cannot tell from the values how much bigger, say, a medium is than a large.
- The central tendency of an ordinal attribute can be represented by its ***mode*** and its ***median*** (middle value in an ordered sequence), but the ***mean*** cannot be defined.

Interval Attributes



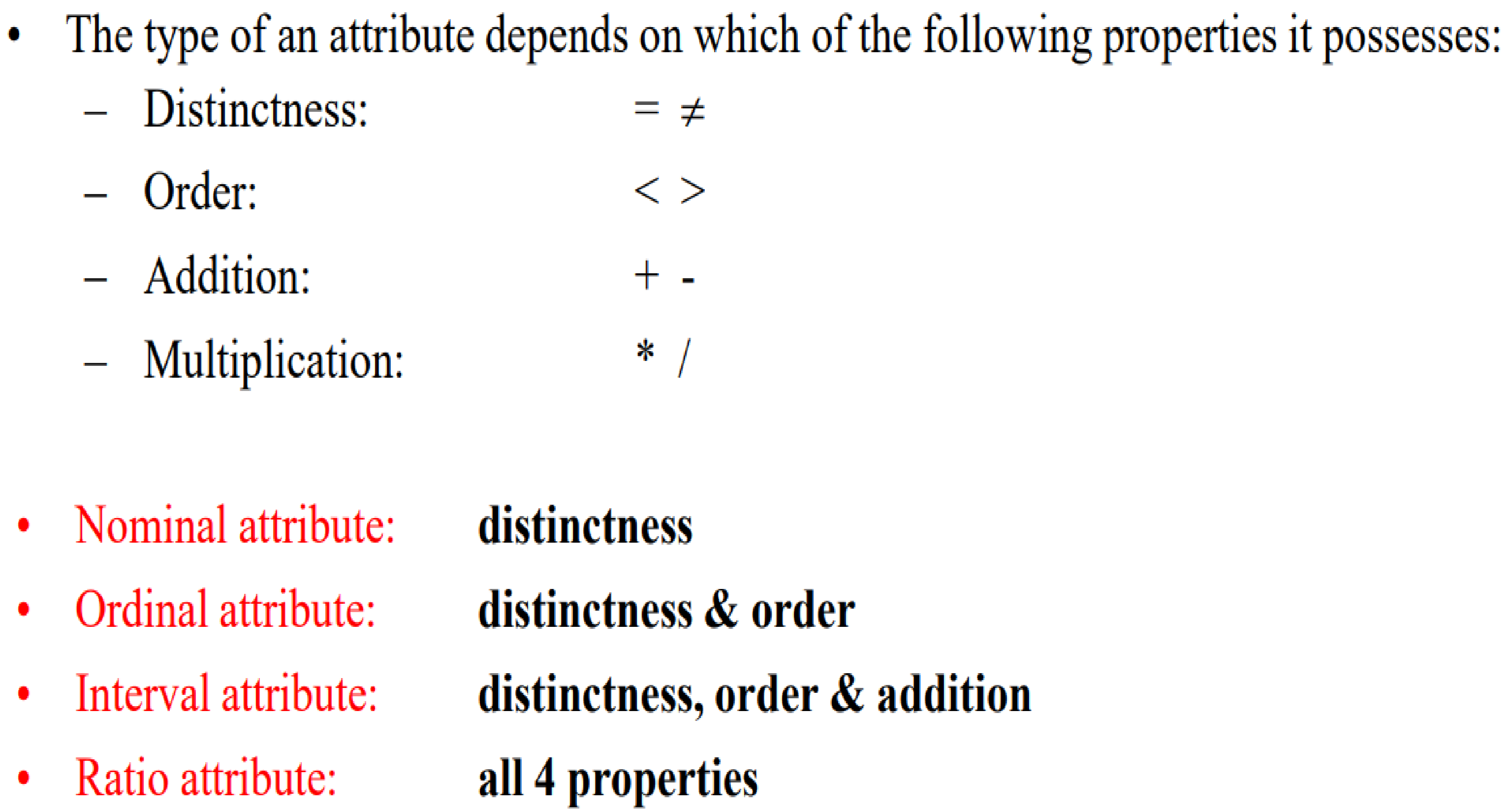
- **Interval attributes** are measured on a ***scale of equal-size units***.
  - We can compare and quantify the difference between values of interval attributes.
- Example: A ***temperature*** attribute is an interval attribute.
  - We can quantify the difference between values. For example, a temperature of 20°C is five degrees higher than a temperature of 15°C.
  - Temperatures in Celsius do not have a **true zero-point**, that is, 0°C does not indicate “no temperature.”
  - Although we can compute the difference between temperature values, we cannot talk of one temperature value as being a multiple of another.
    - Without a true zero, we cannot say, for instance, that 10°C is twice as warm as 5°C . That is, we cannot speak of the values in terms of ratios.
- The central tendency of an interval attribute can be represented by its ***mode***, its ***median*** (middle value in an ordered sequence), and its ***mean***.

Ratio Attributes



- A **ratio attribute** is a numeric attribute with an *inherent zero-point.*
- Example: A ***number_of_words*** attribute is a ratio attribute.
  - If a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value.
- The central tendency of an ratio attribute can be represented by its ***mode***, its ***median*** (middle value in an ordered sequence), and its ***mean***.

Properties of Attribute Values

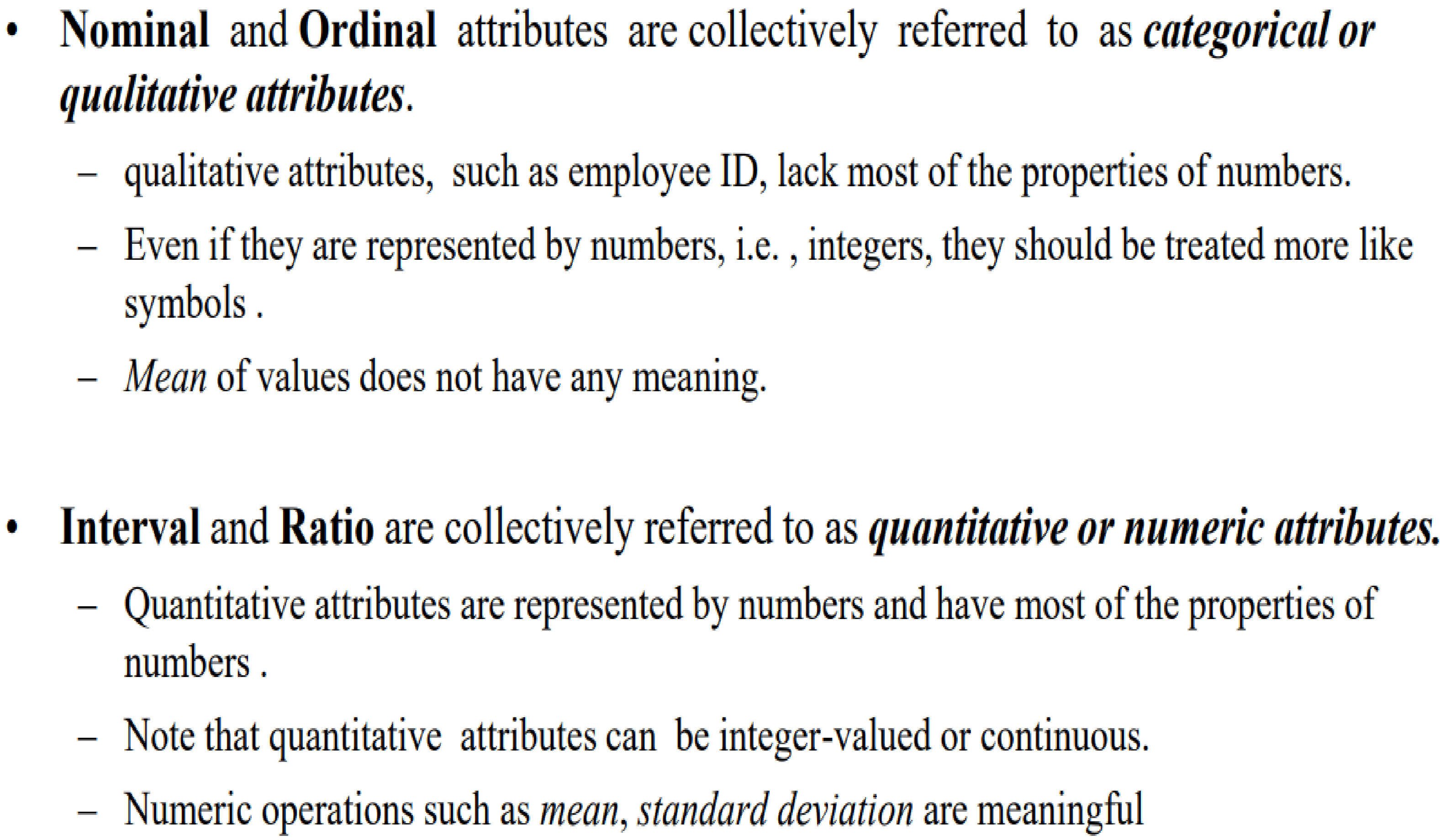


- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

- Nominal attribute: **distinctness**
- Ordinal attribute: **distinctness & order**
- Interval attribute: **distinctness, order & addition**
- Ratio attribute: **all 4 properties**

Attribute Type	Description	Examples
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length,

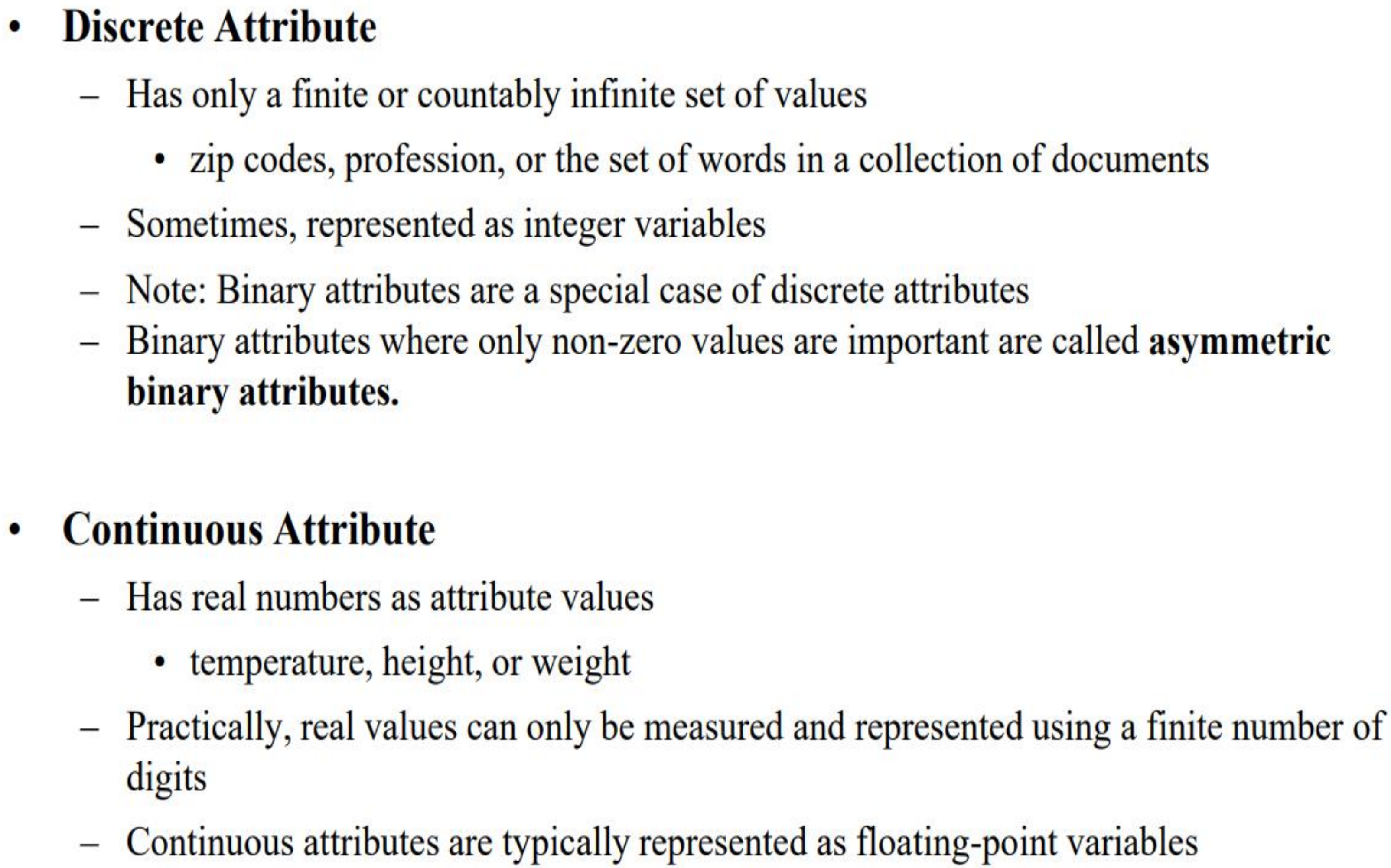
Attribute Types
Categorical (Qualitative) and Numeric (Quantitative)



- **Nominal** and **Ordinal** attributes are collectively referred to as ***categorical or qualitative attributes***.
  - qualitative attributes, such as employee ID, lack most of the properties of numbers.
  - Even if they are represented by numbers, i.e. , integers, they should be treated more like symbols .
  - *Mean* of values does not have any meaning.

- **Interval** and **Ratio** are collectively referred to as ***quantitative or numeric attributes.***
  - Quantitative attributes are represented by numbers and have most of the properties of numbers .
  - Note that quantitative attributes can be integer-valued or continuous.
  - Numeric operations such as *mean*, *standard deviation* are meaningful

Discrete and continuous attributes



- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
  - Binary attributes where only non-zero values are important are called **asymmetric binary attributes.**

- **Continuous Attribute**
  - Has real numbers as attribute values
    - temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Basic Statistical Descriptions of Data

- **Basic statistical descriptions** can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

- For data preprocessing tasks, we want to learn about data characteristics regarding both **central tendency** and **dispersion of the data**.
- **Measures of central tendency** include **mean**, **median**, **mode**, and **midrange**.
- **Measures of data dispersion** include **quartiles**, **interquartile range (IQR)**, and **variance**.

- These descriptive statistics are of great help in understanding the distribution of the data.

Measuring central tendency

- Often, one number can tell us enough about a distribution. This is typically a number that points to the “center” of a distribution. In other words, we can calculate where the “center” of a frequency distribution lies, which is also known as the **central tendency**.

- There are three measures commonly used:

1. Mean
2. median, and
3. mode.

Mean

- The most common and most effective numerical measure of the “***center***” of a set of data is the **arithmetic mean**.

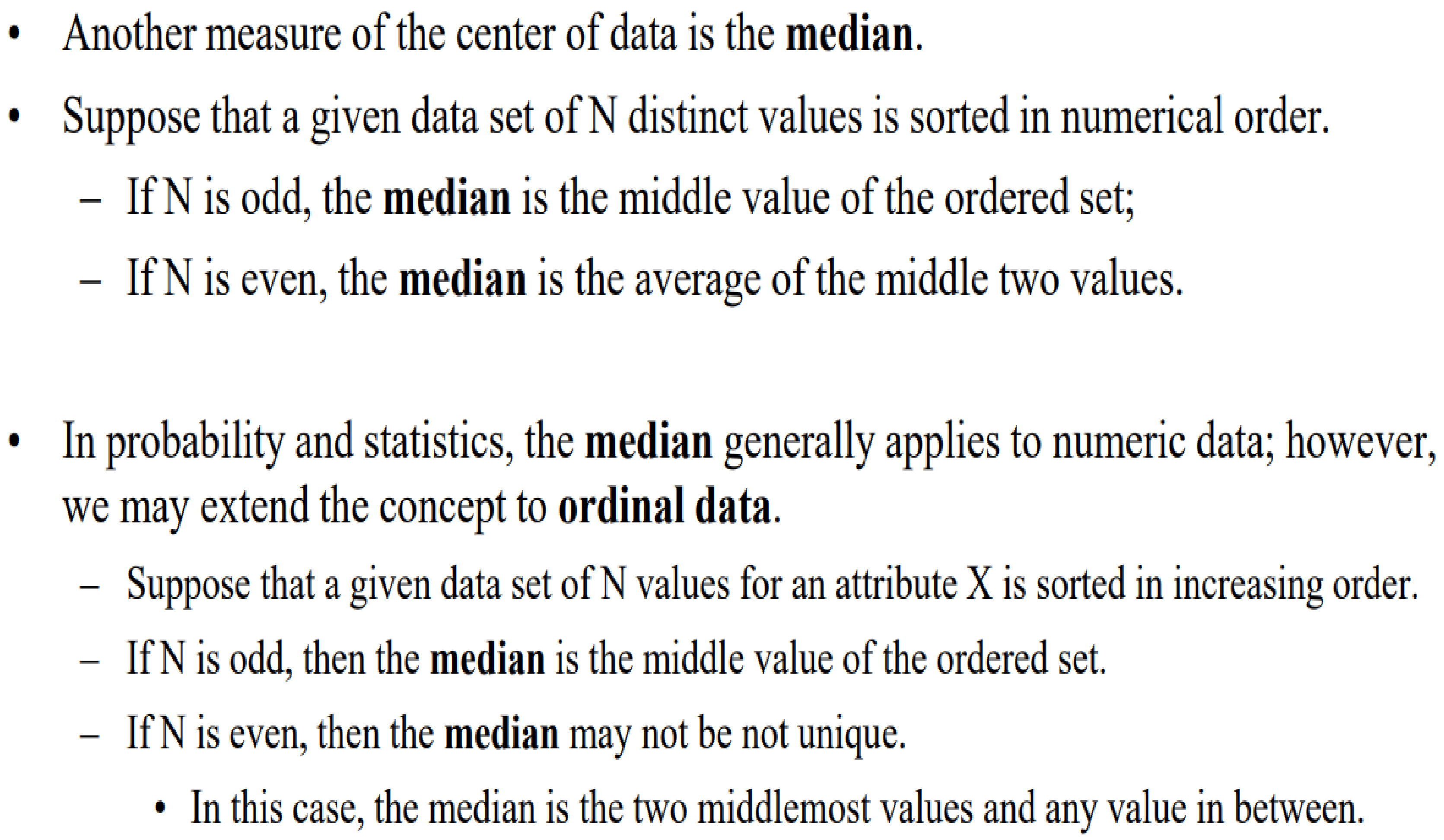
Arithmetic Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- Sometimes, each value x_i in a set may be associated with a weight w_i.
  - The weights reflect the significance and importance attached to their respective values.

Weighted Arithmetic Mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

- Although the mean is the single most useful quantity for describing a data set, it is not always the best way of measuring the center of the data.
  - A major problem with the mean is its sensitivity to extreme (outlier) values.
  - Even a small number of extreme values can corrupt the mean.
- To offset the effect caused by a small number of extreme values, we can instead use the **trimmed mean**,
- **Trimmed mean** can be obtained after chopping off values at the high and low extremes.

Median



- Another measure of the center of data is the **median**.
- Suppose that a given data set of N distinct values is sorted in numerical order.
  - If N is odd, the **median** is the middle value of the ordered set;
  - If N is even, the **median** is the average of the middle two values.

- In probability and statistics, the **median** generally applies to numeric data; however, we may extend the concept to **ordinal data**.
  - Suppose that a given data set of N values for an attribute X is sorted in increasing order.
  - If N is odd, then the **median** is the middle value of the ordered set.
  - If N is even, then the **median** may not be not unique.
    - In this case, the median is the two middlemost values and any value in between.

Mode

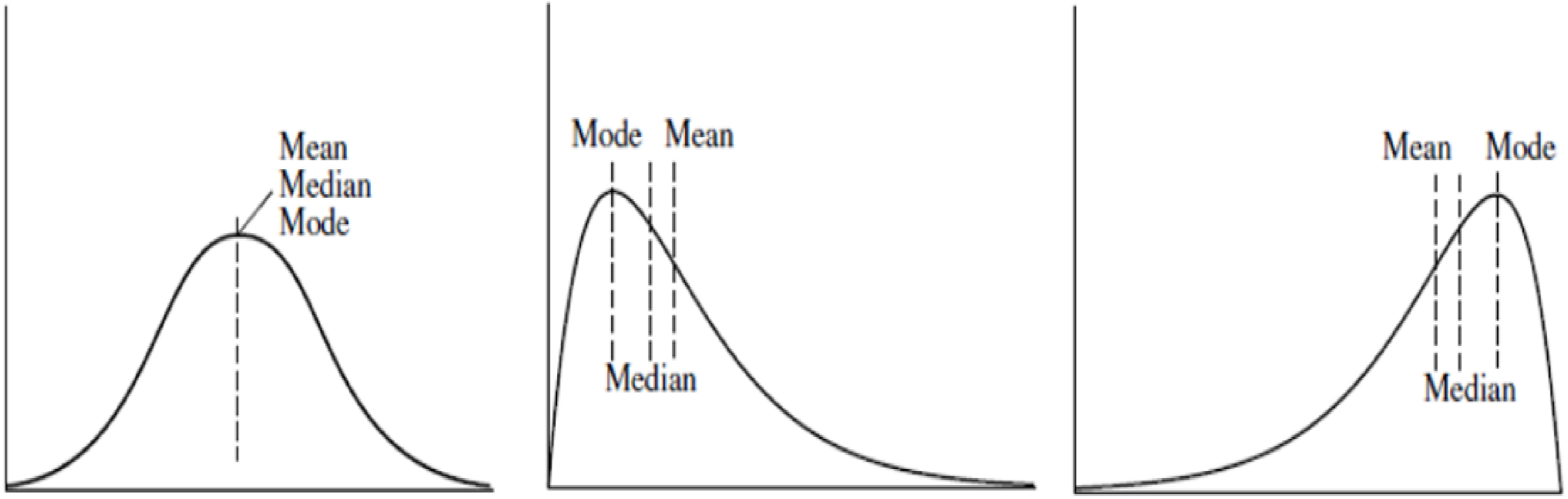
- Another measure of central tendency is the **mode**.

- The **mode** for a set of data is the value that occurs most frequently in the set.
  - It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.
  - Data sets with one, two, or three modes: called ***unimodal***, ***bimodal***, and ***trimodal***.
  - At the other extreme, if each data value occurs only once, then there is no mode.

- *Central Tendency Measures for Numerical Attributes:* ***Mean, Median, Mode***

- *Central Tendency Measures for Categorical Attributes:* ***Mode*** *(Median?)*
  - *Central Tendency Measures for Nominal Attributes:* ***Mode***
  - *Central Tendency Measures for Ordinal Attributes:* ***Mode, Median***

Median, ***mean*** and ***mode*** of *symmetric*, *positively* and *negatively skewed data*



symmetric data *positively skewed data* *negatively skewed data*

- **Positive Skew:** Mean > Median > Mode
- **Negative Skew:** Mean < Median < Mode

Example

What are central tendency measures (mean, median, mode)for the following attributes?

attr1 = {2,4,4,6,8,24}

mean = (2+4+4+6+8+24)/6 = 8	average of all values
median = (4+6)/2 = 5	avg. of two middle values
mode = 4	most frequent item

attr2 = {2,4,7,10,12}

mean = (2+4+7+10+12)/5 = 7	average of all values
median = 7	middle value
mode = any of them (no mode)	all of them has same freq.

attr3 = {xs,s,s,s,m,m,l}

mean is meaningless for categorical attributes.

median = s	middle value
mode = s	most frequent item

Measuring dispersion

- **The degree to which numerical data tend to spread is called the dispersion, or variance of the data.**

The most common ***measures of data dispersion:***

- **Range:** Difference between the largest and smallest values.
- **Interquartile Range (IQR):** range of middle 50%
  - **quartiles**: Q1 (25th percentile), Q3 (75th percentile) IQR=Q3-Q1
  - **five number summary**: Minimum, Q1, Median, Q3, Maximum
- **Variance and Standard Deviation:** *(sample: s, population: σ)*
  - **variance** of N observations:

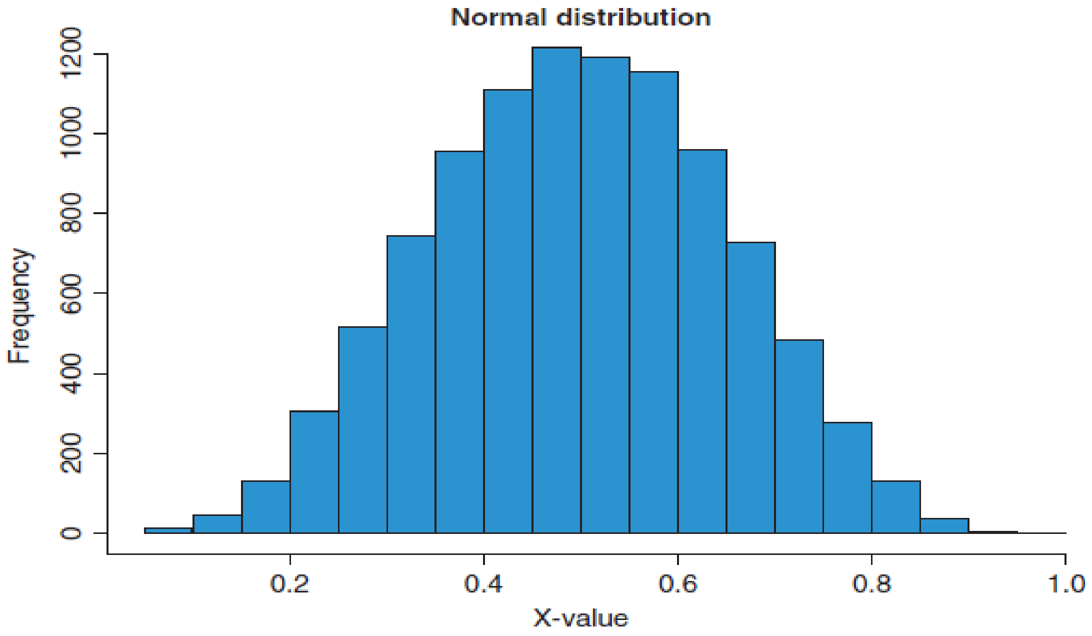
$$\sigma^2 = \frac{1}{n}\sum_{1}^{n}(x_i - \mu)^2 \qquad s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \mu)^2$$

where μ is the mean value of the observations

  - **standard deviation** *σ (s)* is the square root of variance σ^2 (s^2)

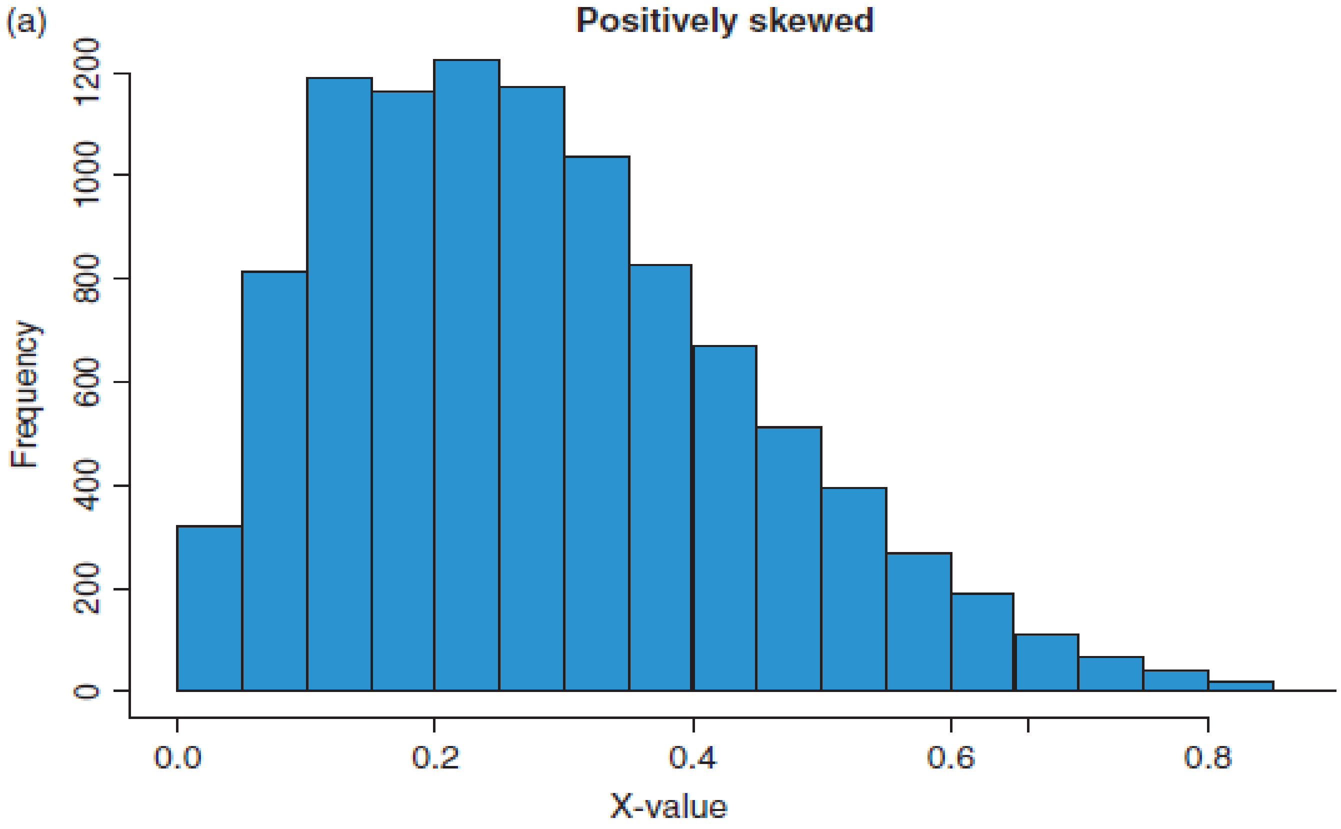
Skewness and kurtosis

- We will often be working with data that are numerical and we will need to understand how those numbers are spread. For that, we can look at the nature of that distribution.
- It turns out that, if the data is normally distributed, various forms of analyses become easy and straightforward.
- **Normal Distribution**. In an ideal world, data would be distributed symmetrically around the center of all scores. Thus, if we drew a vertical line through the center of a distribution, both sides should look the same. This so-called normal distribution, is characterized by a bell-shaped curve, an example of which is shown in Figure

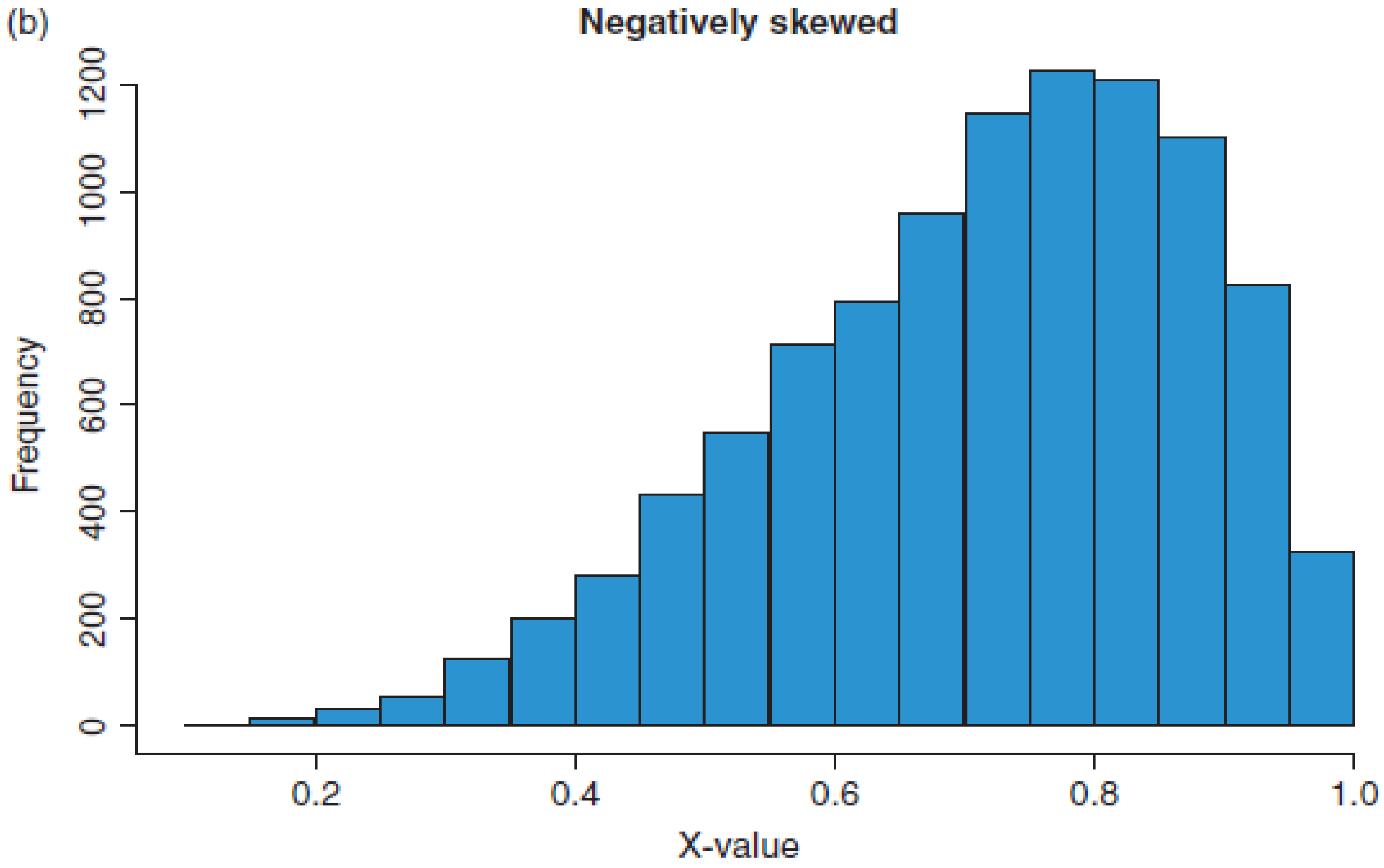


- There are two ways in which a distribution can deviate from normal:
  - Lack of symmetry (called skew)
  - Pointiness (called kurtosis)

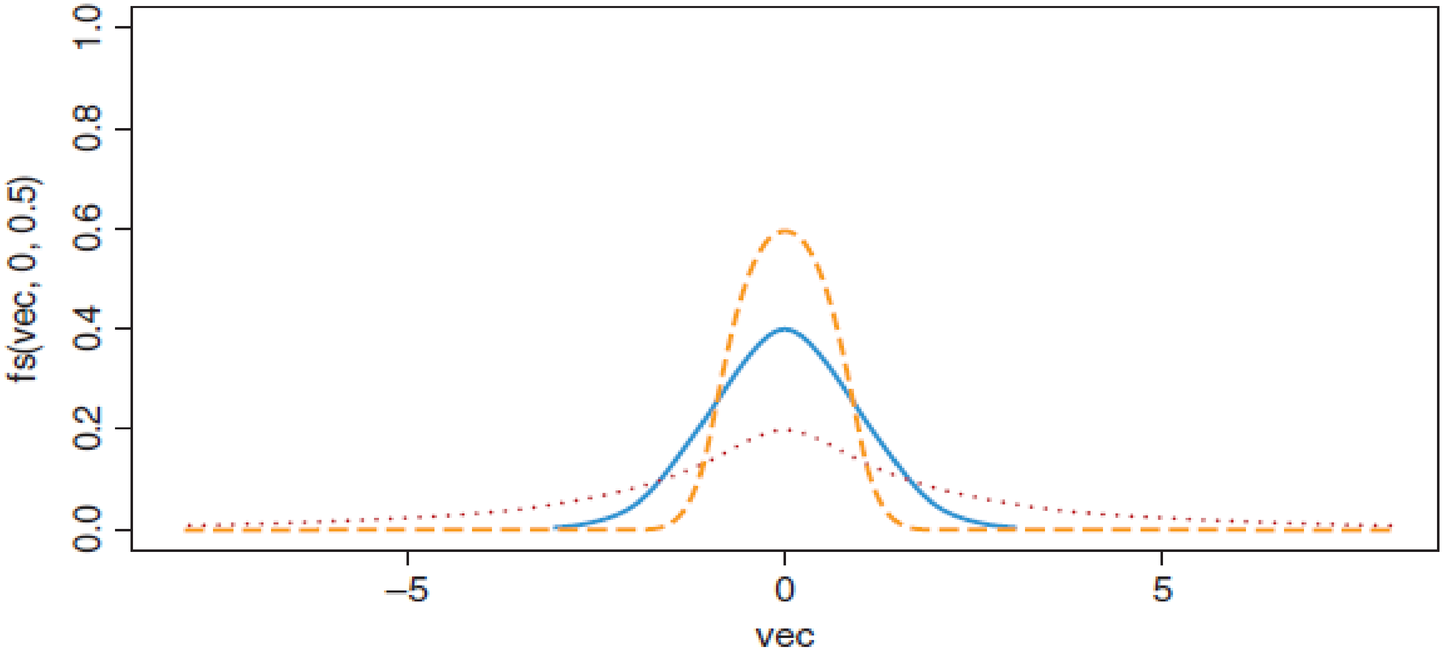
Positive Skew (Right Skew):



Negative Skew (Left Skew):



- **Kurtosis**, on the other hand, refers to the degree to which scores cluster at the end of a distribution (platykurtic) and how “pointy” a distribution is (leptokurtic), as shown in Figure



Examples of different kurtosis in a distribution (orange dashed line represents leptokurtic, blue solid line represents the normal distribution, and red dotted line represents platykurtic).

Understanding relationships using covariance and correlation coefficients

- Covariance and Correlation are two similar measures for assessing how much two attributes change together.

Covariance:

- Covariance is a statistical measure that indicates the extent(direction) to which two variables change together.
- It shows whether the variables tend to move in the same direction (positive covariance), opposite directions (negative covariance), or have no linear relationship (zero covariance).

Example:

- If the covariance between hours studied and exam grades is positive, it means that generally, students who studied more hours tend to have higher grades.

- $\mathrm{Cov}(X, Y) = \Sigma[(X - \mu_x)(Y - \mu_\gamma)] / N$

where:

  - X and Y are the two variables
  - μ_x and μ_γ are the means of X and Y, respectively
  - N is the number of data points

- The value of covariance lies between -∞ and +∞.

Example: Calculate the coefficient of covariance for the following data:

X	2	8	18	20	28	30
Y	5	12	18	23	45	50

Number of observations = 6

Mean of X = 17.67

Mean of Y = 25.5

Cov(X, Y)

$= (\frac{1}{6}) [(2 - 17.67)(5 - 25.5) + (8 - 17.67)(12 - 25.5) + (18 - 17.67)(18 - 25.5) + (20 - 17.67)(23 - 25.5) + (28 - 17.67)(45 - 25.5) + (30 - 17.67)(50 - 25.5)]$

$= 157.83$

Correlations

- Correlation is a statistical analysis that is used to measure and describe the strength and direction of the relationship between two variables.
  - Strength indicates how closely two variables are related to each other,
  - direction indicates how one variable would change its value as the value of the other variable changes.

- Correlation is a simple statistical measure that examines how two variables change together over time.
- for example, “umbrella” and “rain.”
- If someone who grew up in a place where it never rained saw rain for the first time, this person would observe that, whenever it rains, people use umbrellas. They may also notice that, on dry days, folks do not carry umbrellas.

- By definition, “rain” and “umbrella” are said to be correlated!

- An important statistic, the **Pearson’s r correlation**, is widely used to measure the degree of the relationship between linear related(continuous) variables.
- The following formula is used to calculate the Pearson’s r correlation:

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}},$$

where

r = Pearson’s r correlation coefficient,
N = number of values in each dataset,
$\sum xy$ = sum of the products of paired scores,
$\sum x$ = sum of x scores,
$\sum y$ = sum of y scores,
Σx2 = sum of squared x scores, and
Σy2 = sum of squared y scores.6

The value of correlation lies between -1 and +1.

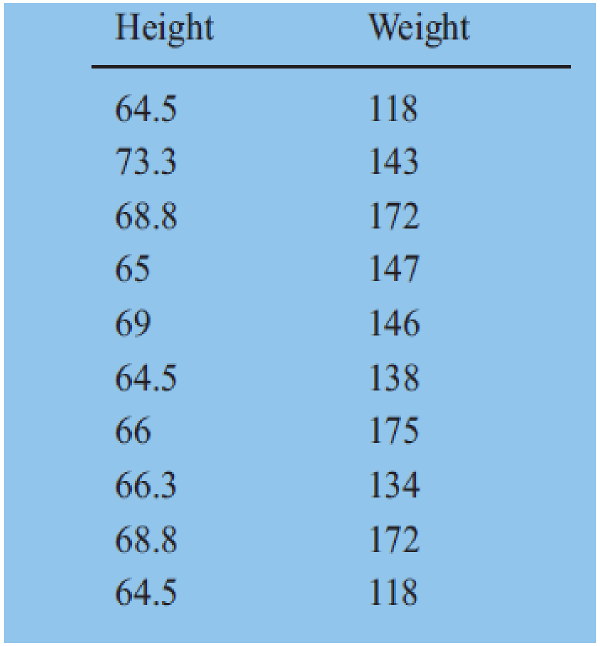
Limitations:

•Only measures linear relationships.

•Sensitive to outliers.

Example:

- Let us use the formula and calculate Pearson’s r correlation coefficient for the height– weight pair with the data provided.



Height	Weight
64.5	118
73.3	143
68.8	172
65	147
69	146
64.5	138
66	175
66.3	134
68.8	172
64.5	118

N = number of values in each dataset = 10
Σ xy = sum of the products of paired scores = 98,335.30
Σ x = sum of x scores = 670.70
Σ y = sum of y scores = 1463
Σ x^2 = sum of squared x scores = 45,058.21
Σ y^2 = sum of squared y scores = 218,015

The Pearson’s r correlation formula gives us 0.39 (approximated to two decimal places) as the correlation coefficient.
This indicates two things: (1) “height” and “weight” are positively related, which means that, as one goes up, so does the other; and (2) the strength of their relation is medium.

Spearman's rank correlation coefficient

- Used to measure The strength and direction of the **monotonic** relationship between two variables.

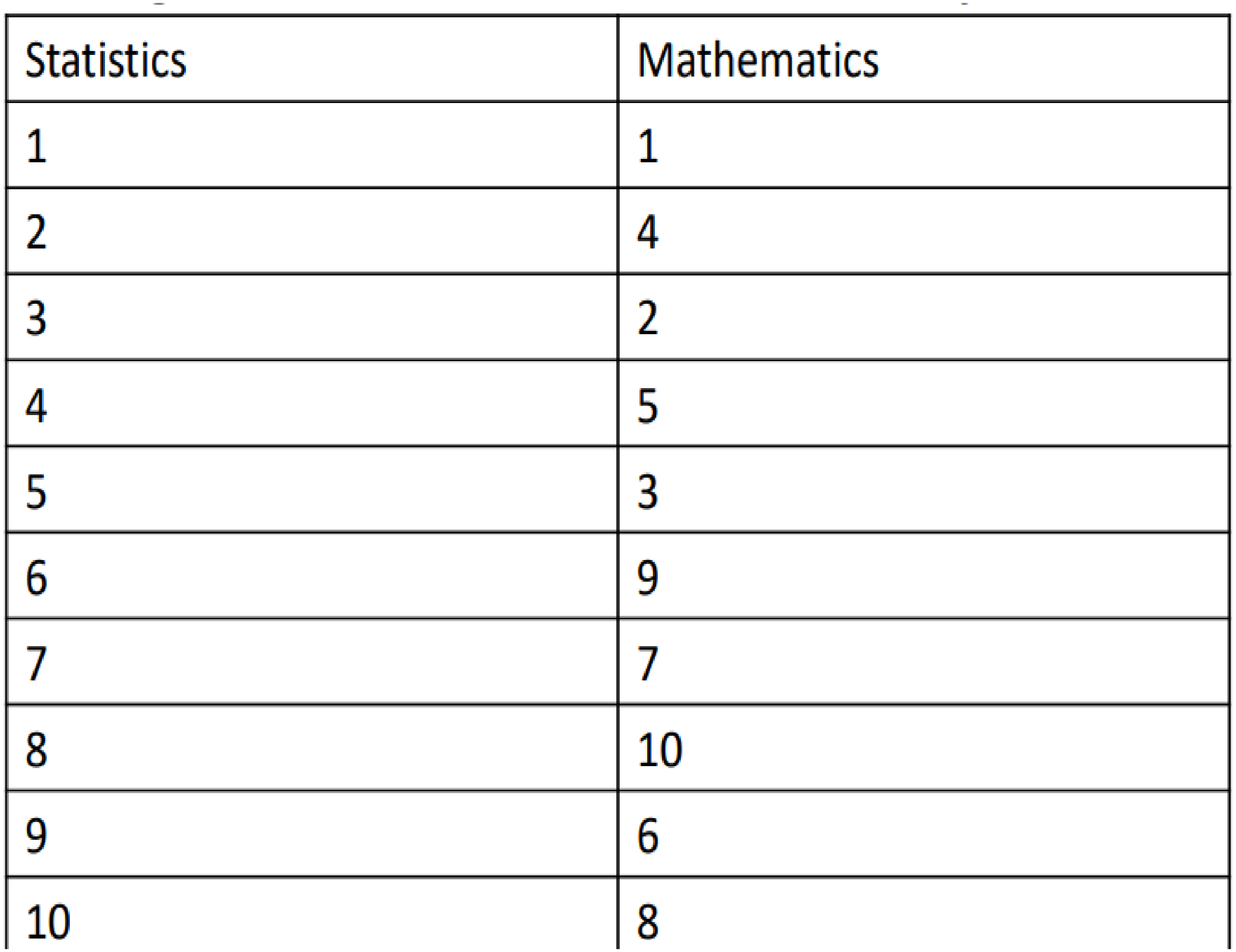
Monotonic relationship: The variables tend to change together, but not necessarily at a constant rate. It could be an increasing or decreasing trend.

The Spearman coefficient is denoted with the Greek letter rho (ρ).

$$\rho = 1 - \{6\sum d^2 / [n(n2-1)]\}$$

- ρ (rho) represents the Spearman's rank correlation coefficient.
- d is the difference in ranks for each pair of data points.
- n is the number of data points

Example: The following are the ranks obtained by 10 students in Statistics & Mathematics subject. To what extent is the knowledge of the students in the two subjects are related?



Statistics	Mathematics
1	1
2	4
3	2
4	5
5	3
6	9
7	7
8	10
9	6
10	8

Statistics (X)	Mathematics (Y)	d = X-Y	d^2
1	1	0	0
2	4	-2	4
3	2	1	1
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4
			$\sum d^2 = 36$

$\rho = 1 – \{6\sum d^2 / [n(n2−1)]\}$

On substituting the values,

$\rho = 1 – \{(6*36)/ [10(100−1)]\}$

$\rho = 1 – 0.2182$ So,

$\rho = 0.7818$

Indicating a strong positive monotonic relationship between variables X and Y. This means that as the values of X increase, the values of Y tend to increase as well.

- The value of correlation lies between -1 and +1.

Limitations:

- When the data is not normally distributed.
- When dealing with ordinal data (e.g., rankings).
- When the relationship is not necessarily linear

Collecting samples

- Statistical analysis often requires data collected from samples rather than entire populations due to time, cost, and feasibility constraints.
- Since analyzing an entire population is impractical, we collect **samples** that represent the population. The quality of inferences depends on how well the sample reflects the population.
- Sampling is the process of selecting a subset of individuals or objects from a larger population to study and make inferences about the entire population.

- **Cost-effectiveness:** Studying the entire population can be expensive and time-consuming.
- **Feasibility:** It may be impossible to study every member of a large population.
- **Practicality:** In some cases, the entire population may not be accessible.

Types of Sampling Methods

Simple Random Sampling: Every member of the population has an equal chance of being selected.

Example: Drawing 100 students' names randomly from a university database.

Stratified Sampling: The population is divided into subgroups (strata), and a random sample is taken from each stratum.

Example: Selecting equal proportions of students from different grade levels.

Systematic Sampling: Selecting every kkkth individual from a list.

Example: Choosing every 10th customer from a store's entry list.

Cluster Sampling: The population is divided into clusters, and a random sample of clusters is selected.

Example: Selecting schools randomly and surveying all students in chosen schools.

Convenience Sampling: Selecting individuals who are easily accessible

Example: Surveying people at a shopping mall about their opinions on a new product.

Performing parametric tests

- Parametric tests are statistical tests that make assumptions about the distribution of the population from which the sample is drawn.

Typical assumptions are:

- Normality: Data have a normal distribution (or at least is symmetric) (bell-shaped curve)
- Homogeneity of variances: Data from multiple groups have the same variance
- Linearity: Data have a linear relationship
- Independence: Data are independent

- A **hypothesis** is a testable prediction or educated guess about the relationship between two or more variables. It's a crucial step in the scientific method.

Example: Smokers are at higher risk of developing lung cancer than non-smokers.

Key Characteristics:

- **Testable:** The hypothesis must be able to be tested through experimentation or observation.
- **Falsifiable:** It must be possible to prove the hypothesis wrong.
- **Specific:** The hypothesis should be clear, concise, and avoid ambiguity.
- **Measurable:** The variables involved should be measurable and quantifiable.

Types of Hypotheses:

Null Hypothesis (H0):

States that there is no significant difference or relationship between the variables. It's the default assumption.

Example: There is no difference in the average height between men and women.

Alternative Hypothesis (H1 or Ha):

States that there is a significant difference or relationship between the variables.

It can be:

Directional: Specifies the direction of the difference (e.g., "Men are taller than women").

Non-directional: States that there is a difference, but does not specify the direction (e.g., "There is a difference in height between men and women").

Example: Does studying for more hours improve exam scores?

- **Null Hypothesis (H0):** There is no relationship between the number of hours studied and exam scores.
- **Alternative Hypothesis (H1):** Students who study for more hours will have higher exam scores.

Z-test

- z-test is a statistical method for the comparison of mean in a sample from the normally distributed population or between two independent samples.
- Statistical test to validate the hypothesis (accept or reject) when the data is normally distributed.

z-test is used when:

- population standard deviation is known.
- Sample size is greater than 30

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

x̄ : mean of the sample.

μ : mean of the population.

σ : Standard deviation of the population.

Example: A professor claims that the **average score** of students in a statistics exam is **75**. We collect a sample of 40 students with an average score of **72** and standard deviation **8**. Can we conclude the students' performance is different?

Population mean (μ) = **75** (claimed mean)
Sample mean ($\bar{x}$) = **72**
Standard deviation (σ) = **8**
Sample size (n) = **40**
Significance level (α) = **0.05** (if not given, we assume the common 5%)

- **Formulate Hypotheses**
- **Null Hypothesis (H_0)**: The average score is 75

$$H_0: \mu=75$$

- **Alternative Hypothesis (H_1)**: The average score is different from 75

$$H_1:\mu \neq 75$$

This is a **two-tailed** test since we are testing if the mean is different (not just greater or smaller).

Standard Error (SE) = σ / √n

- Since the population standard deviation (σ) is unknown, we'll use the sample standard deviation (s) as an estimate. This is acceptable because we have a large sample size (n = 40).

$$SE = s / \sqrt{n} = 8 / \sqrt{40} \approx 1.26$$

Calculate the z-score

$$z = (\bar{x} - \mu) / SE$$

$$z = (72 - 75) / 1.26$$

$$z = -3 / 1.26$$

$$z \approx -2.38$$

Determine the Critical Value

We'll use a significance level (α) of 0.05 for a two-tailed test.

For a two-tailed test at α = 0.05, the critical z-values are ±1.96.

Compare the Calculated z-score to the Critical Value

Our calculated z-score (-2.38) is less than the lower critical value (-1.96).

Since our calculated z-score falls in the rejection region, we **reject the null hypothesis**.

t - test

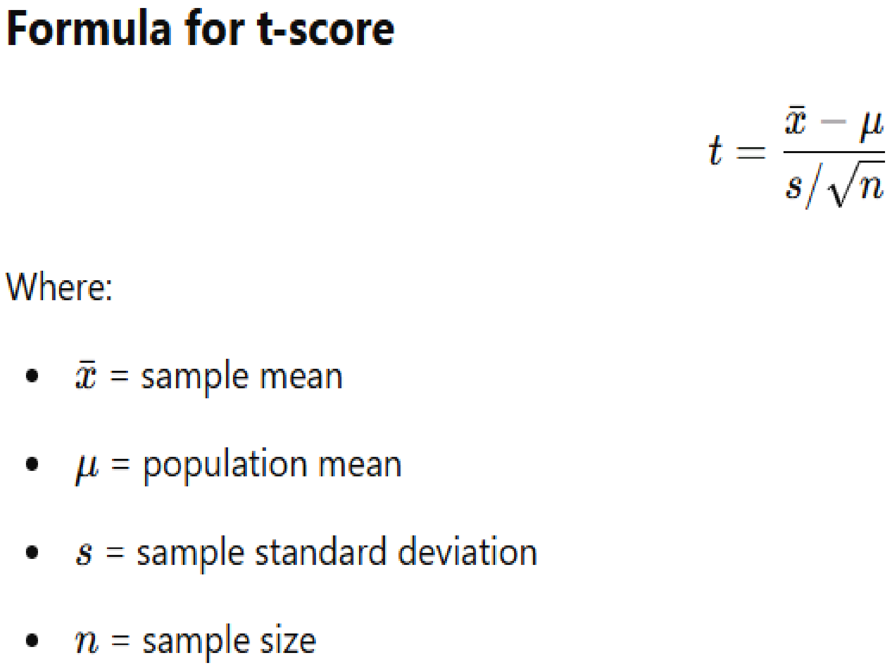
- The **t-test** is a statistical test procedure that tests whether there is a significant difference between the means of two groups when the **sample size is small (n < 30)** or when the **population standard deviation (σ) is unknown**.
- Assumes the population standard deviation (σ) is unknown and is estimated using the sample standard deviation (s).

•**Independent samples t-test:** Compares the means of two independent groups.
•*Example:* Comparing the average test scores of two different teaching methods.

•**Paired samples t-test:** Compares the means of two related groups (e.g., before-and-after measurements).
•*Example:* Comparing the blood pressure of patients before and after taking a medication.

Types of T-Tests

1. **One-Sample T-Test** → Compares the mean of a sample to a known population mean.
2. Used to check if the mean of a sample is significantly different from a known population mean.



Formula for t-score

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

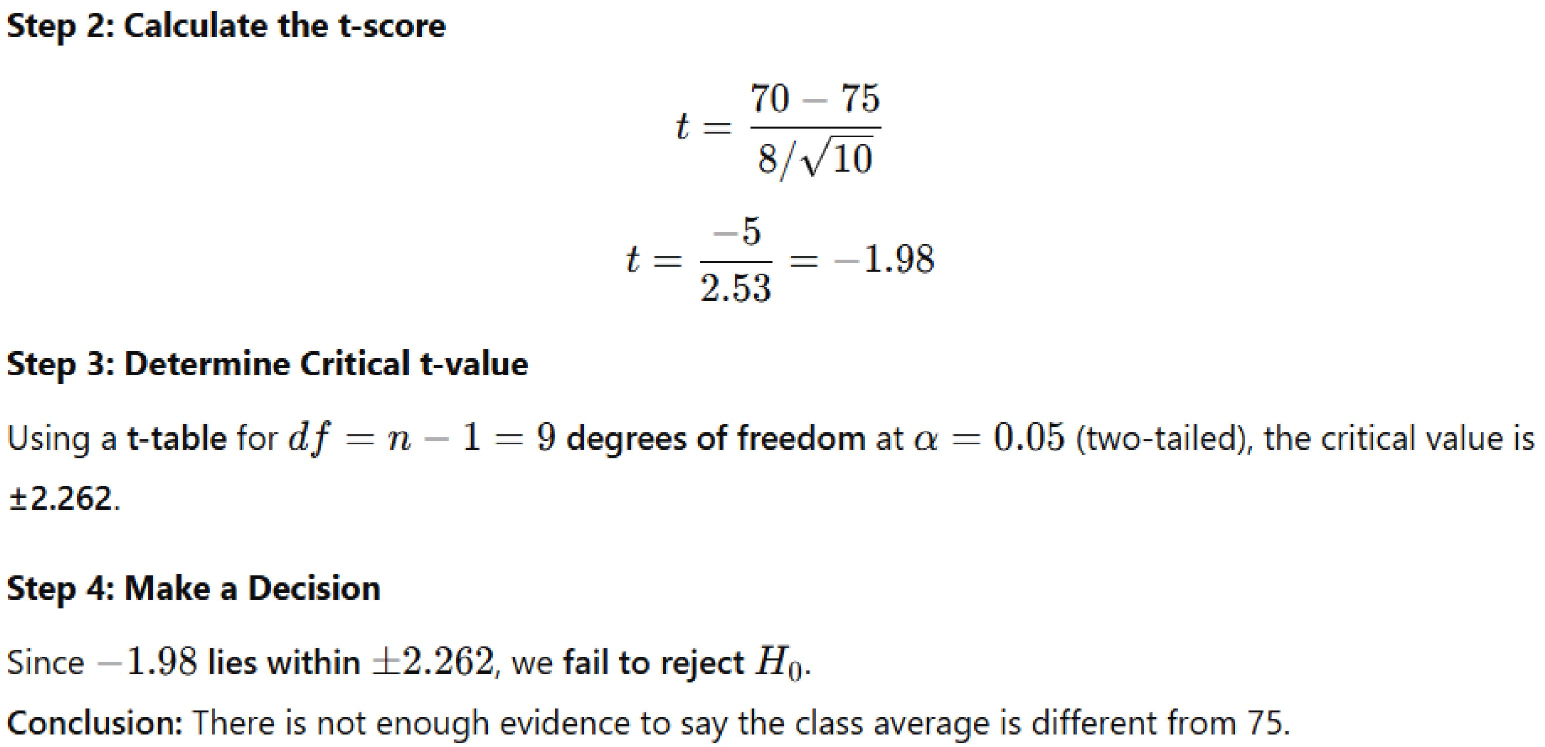
Where:

- $\bar{x}$ = sample mean
- μ = population mean
- s = sample standard deviation
- n = sample size

- A teacher claims that the average score in a class is **75**. A sample of **10 students** has an average score of **70** with a standard deviation of **8**. Is the claim correct at a **5% significance level**?

Step1: Define Hypotheses

- **Null Hypothesis (H0)**: The sample mean is equal to the population mean. H0:μ=75
- **Alternative Hypothesis (H1)**: The sample mean is significantly different. H1:μ≠75



Step 2: Calculate the t-score

$$t = \frac{70 - 75}{8/\sqrt{10}}$$

$$t = \frac{-5}{2.53} = -1.98$$

Step 3: Determine Critical t-value

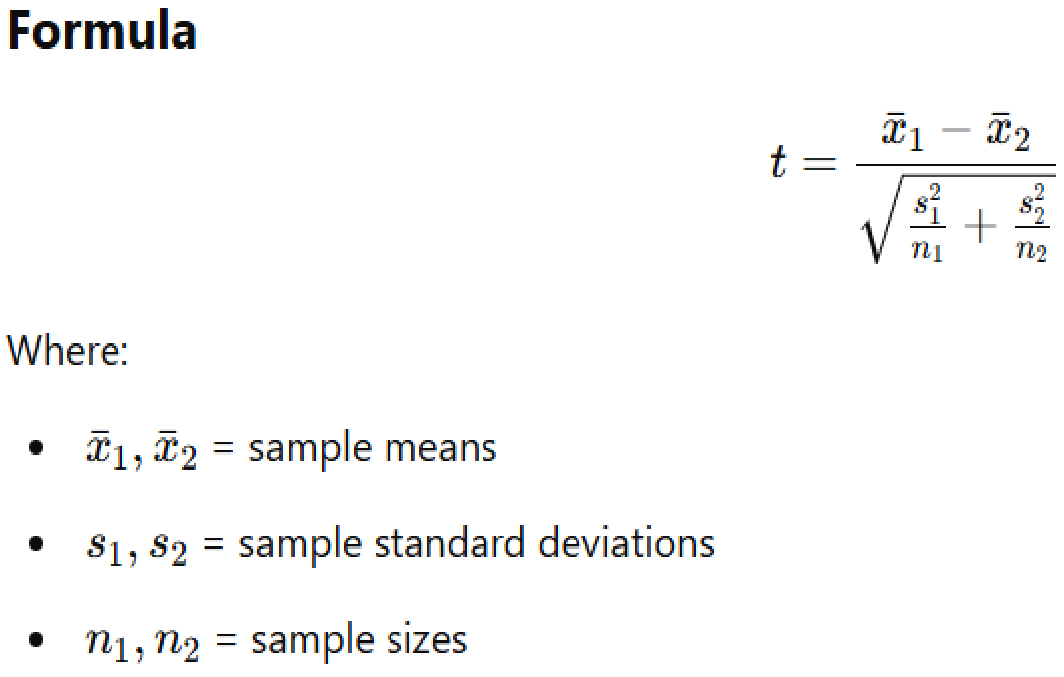
Using a **t-table** for $df = n - 1 = 9$ **degrees of freedom** at $\alpha = 0.05$ (two-tailed), the critical value is **±2.262**.

Step 4: Make a Decision

Since -1.98 **lies within** ± 2.262, we **fail to reject** H_0.

Conclusion: There is not enough evidence to say the class average is different from 75.

- **Two-Sample T-Test (Independent T-Test)** → Compares the means of two independent groups.



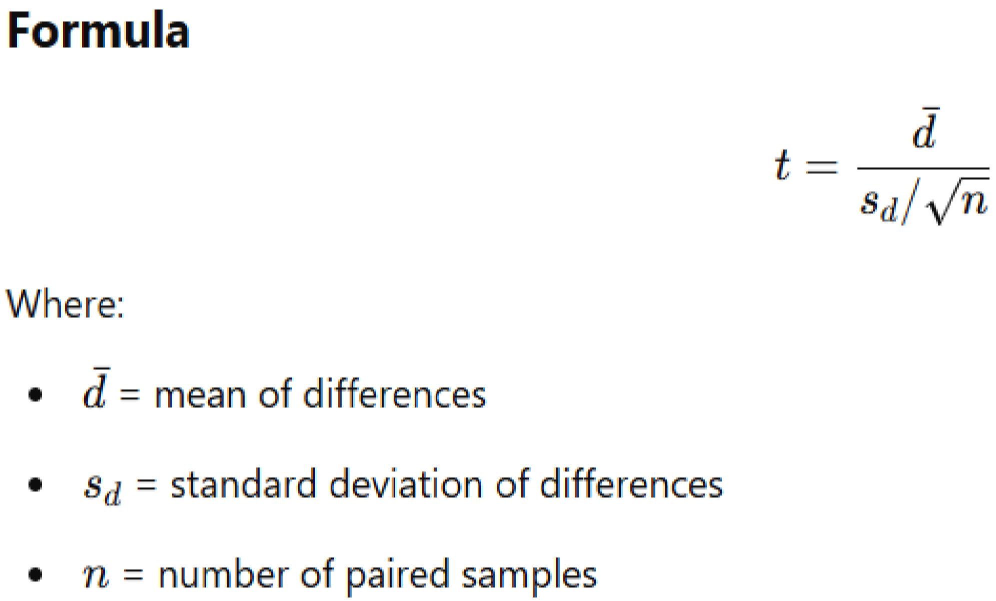
Formula

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\bar{x}_1, \bar{x}_2$ = sample means
- s_1, s_2 = sample standard deviations
- n_1, n_2 = sample sizes

- **Paired T-Test (Dependent T-Test)** → Compares the means of two related groups (before and after measurements).



Formula

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

Where:

- $\bar{d}$ = mean of differences
- s_d = standard deviation of differences
- n = number of paired samples

ANOVA

•When we have **more than two groups**, performing multiple **t-tests** increases the probability of making a **Type I error** (false positive).
•ANOVA helps to **compare multiple groups** in a single test without increasing error rates.

- **ANOVA (Analysis of Variance):** Used to compare the means of three or more groups.
- *Example:* Comparing the average crop yields of three different fertilizers.

Types of ANOVA

- **One-way ANOVA:** Used to compare the means of three or more groups based on one independent variable.
  - **Example:** Comparing the average crop yields of three different fertilizers.
- **Two-way ANOVA:** Used to compare the means of groups based on two independent variables.
  - **Example:** Comparing the average test scores of students based on both their teaching method (e.g., online, in-person, hybrid) and their study habits (high, medium, low).
- **Repeated Measures ANOVA:** Used when the same subjects are measured multiple times under different conditions.
  - **Example:** Measuring the blood pressure of patients before, during, and after a medication treatment.

Descriptive Statistics

- Descriptive statistics help to summarize a provided dataset and identify the most significant features of the data under consideration.

Understanding statistics

- In data science, both qualitative and quantitative analyses are important aspects.
- In particular, the quantitative analysis of any dataset requires an understanding of statistical concepts.
- Statistics is a branch of mathematics that deals with collecting, organizing, and interpreting data.
- Hence, by using statistical concepts, we can understand the nature of the data, a summary of the dataset, and the type of distribution that the data has.

Distribution functions

- Refers to the function that gives the probability of all possible values of a random variable.
- Often called **cumulative distribution function**
- It shows how the probabilities are assigned to the different possible values of the random variable.
- A **Probability Distribution Function (PDF)** is a mathematical function that describes the likelihood of different outcomes in a random experiment. For any random variable **X**, where its value is evaluated at the points ‘x’, then the probability distribution function gives the probability that X takes the value less than equal to x.
- We represent the probability distribution as, $F(x) = P\ (X \leq x)$
- The cumulative probability for a closed interval(a, b] is given by:

$$\mathbf{P(a < X \leq b) = F(b) - F(a)}$$

- **For probability distribution function the value of the variable lies between 0 and 1: $0 \leq F(x) \leq 1$**

Uniform distribution

- A uniform distribution is a probability distribution where all outcomes are equally likely.
- **Discrete Uniform Distribution:**
  - Deals with a finite number of equally likely outcomes.
  - **Example:**
    - Rolling a fair die: Each number (1-6) has an equal probability of 1/6.
    - Drawing a card from a well-shuffled deck: Each card has an equal probability of being drawn.
- **Continuous Uniform Distribution:**
  - Deals with outcomes that can take on any value within a specified range.
  - **Example:**
    - Spinning a spinner that lands on a number between 0 and 1 with equal probability for any point within that range.
    - The arrival time of a bus, assuming it arrives randomly within a 10-minute window.

- The probability density function(PDF) of the continuous uniform distribution is:

  $f(x) = 1 / (b - a)$ for $a \leq x \leq b$

  $f(x) = 0$ otherwise

Example: Let's say a bus arrives at a bus stop randomly between 9:00 AM and 9:15 AM. We can model this situation using a continuous uniform distribution.

- **Range:** The possible arrival times range from 9:00 AM (0 minutes) to 9:15 AM (15 minutes).
- **Distribution:** The probability of the bus arriving at any specific time within this 15-minute window is equal.

In our example, a = 0 minutes (9:00 AM) and b = 15 minutes (9:15 AM). So, the PDF is:

$f(x) = 1 / (15 - 0) = 1/15$ for $0 \leq x \leq 15$

$f(x) = 0$ otherwise

- The values a and b are the parameters of the uniform distribution. It can be shown that $E(X) = (a + b) / 2$ and $V(X) = (b - a)^2 / 12$

In our bus arrival example:

- **a = 0 minutes** (minimum arrival time)
- **b = 15 minutes** (maximum arrival time)

Mean (μ)

- For a continuous uniform distribution, the mean is given by:

$$\mu = (a + b) / 2$$

- In our case: $\mu = (0 + 15) / 2 = 7.5$ minutes

Variance (σ²)

- For a continuous uniform distribution, the variance is given by:

$$\sigma^2 = (b - a)^2 / 12$$

- In our case: $\sigma^2 = (15 - 0)^2 / 12 = 225 / 12 = 18.75$

Therefore:

- The mean arrival time of the bus is 7.5 minutes (or 9:07:30 AM).
- The variance of the arrival time is 18.75 minutes².

Normal distribution

- In probability theory and statistics, the **Normal Distribution**, also called the **Gaussian Distribution**, is the most significant continuous probability distribution. Sometimes it is also called a bell curve.
- We use this distribution to represent
- We define Normal Distribution as the probability density function of any continuous random variable for any given system. Now for defining Normal Distribution suppose we take f(x) as the probability density function for any random variable X.
- Also, the function is integrated between the interval, (x, {x + dx}) then,

$$f(x) \geq 0 \ \forall \ x \in (-\infty,+\infty),$$

$$\int_{-\infty}^{+\infty} \mathbf{f(x) = 1}$$

- The probability density function of normal or gaussian distribution is given by;

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where,
•x is the variable
•μ is the mean
•σ is the standard deviation

Example

- **Calculate the probability density function of normal distribution using the following data. x = 3, μ = 4 and σ = 2.**

Solution: Given, variable, x = 3

Mean = 4 and

Standard deviation = 2

By the formula of the probability density of normal distribution, we can write;

$$f(3,4,2)=\frac{1}{2\sqrt{2\pi}}e^{\frac{-(3-2)^2}{2\times2^2}}$$

Hence, f(3,4,2) = 1.106.

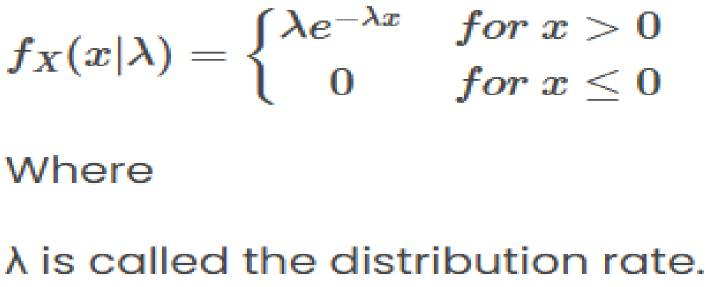
Exponential distribution

- Exponential distributions are a class of continuous probability distribution.
- An exponential distribution arises naturally when modeling the time between independent events that happen at a constant average rate.

- For example, if you receive 3 calls on average between 8am-5pm each day, then the hours you wait for the first call since 8am tomorrow should follow an exponential distribution with parameter

λ = 3 calls per 9 hrs = 1/3.

The average time you wait for the new call since last call is the expectation of the distribution: 1/ λ = 3 hrs . The probability density function is ,



$$f_X(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & for\ x > 0 \\ 0 & for\ x \leq 0 \end{cases}$$

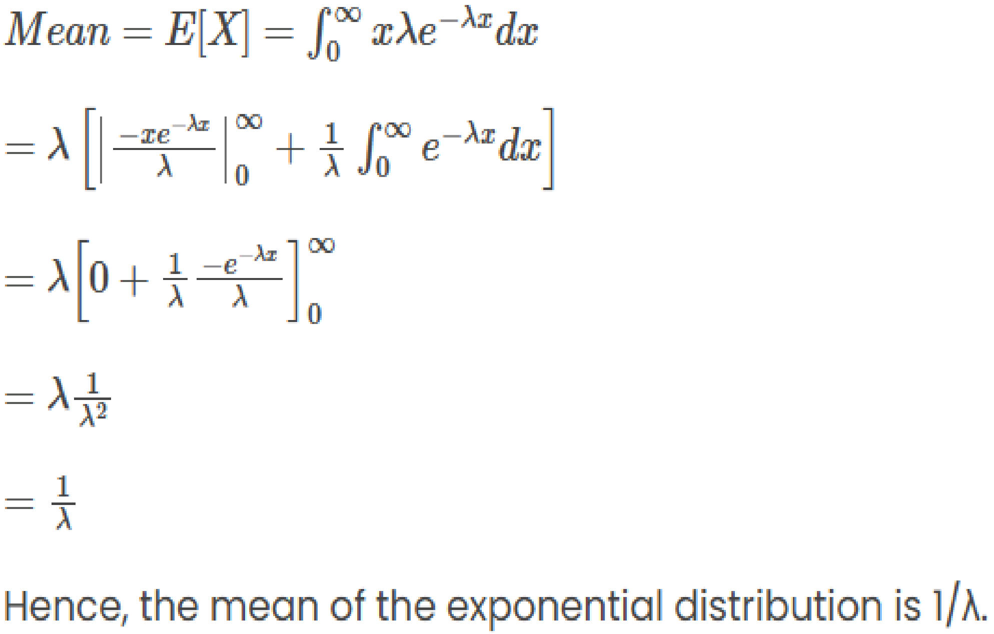
Where

λ is called the distribution rate.

Mean and Variance of Exponential Distribution

Mean:

- The mean of the exponential distribution is calculated using the integration by parts.



$$Mean = E[X] = \int_0^\infty x\lambda e^{-\lambda x} dx$$

$$= \lambda \left[\left| \frac{-xe^{-\lambda x}}{\lambda} \right|_0^\infty + \frac{1}{\lambda} \int_0^\infty e^{-\lambda x} dx \right]$$

$$= \lambda \left[0 + \frac{1}{\lambda} \frac{-e^{-\lambda x}}{\lambda} \right]_0^\infty$$

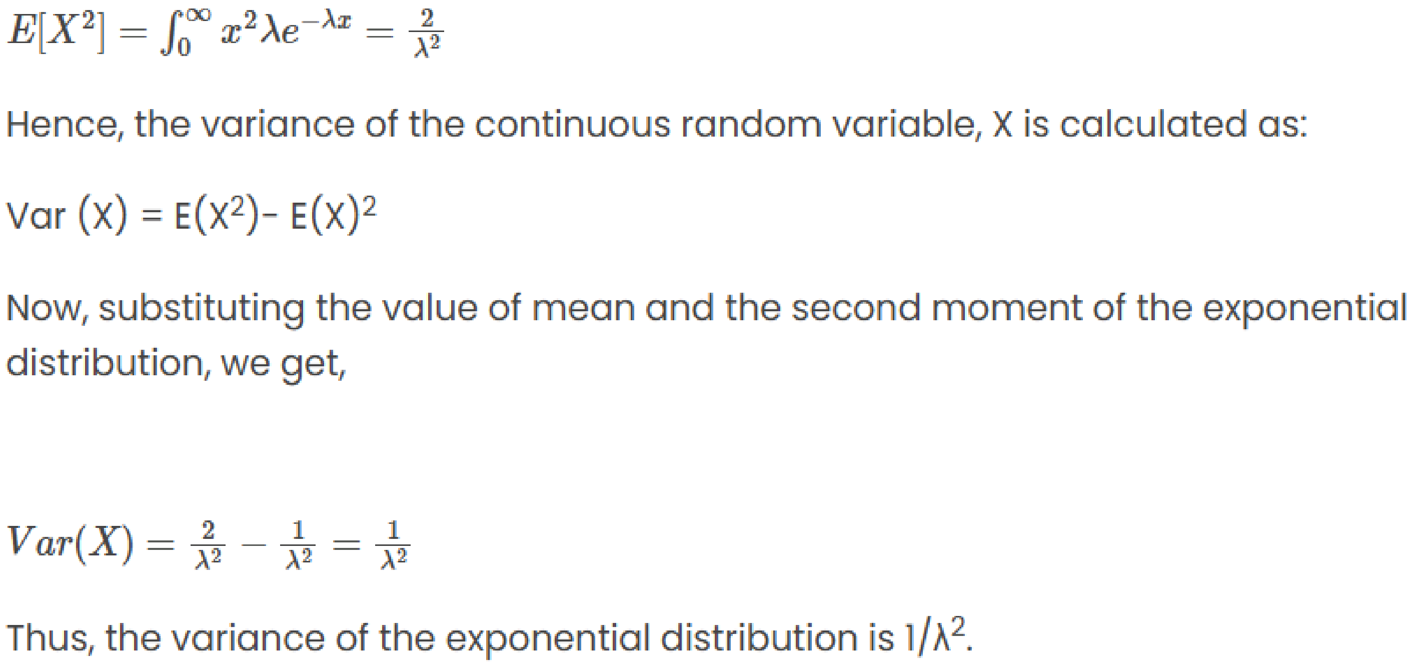
$$= \lambda \frac{1}{\lambda^2}$$

$$= \frac{1}{\lambda}$$

Hence, the mean of the exponential distribution is 1/λ.

Variance:

- To find the variance of the exponential distribution, we need to find the second moment of the exponential distribution, and it is given by:



$$E[X^2] = \int_0^\infty x^2 \lambda e^{-\lambda x} = \frac{2}{\lambda^2}$$

Hence, the variance of the continuous random variable, X is calculated as:

Var (X) = E(X^2)- E(X)^2

Now, substituting the value of mean and the second moment of the exponential distribution, we get,

$$Var(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Thus, the variance of the exponential distribution is 1/λ^2.

Binomial distribution

- Binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent binary (yes/no) experiments, each of which yields success with probability p.
- Such a success/failure experiment is also called a Bernoulli experiment or Bernoulli trial.
- In fact, when n = 1, the binomial distribution is a Bernoulli distribution. Let X be a random variable following a Binomial distribution B(n, p), then for any integer k= 0, 1, 2, …, n, its probability mass function is

$$P(k{:}n,p) = {}^{n}C_{k}\, p^{k}\, (q)^{n-k}$$

Where,

n = the number of experiments

k = 0, 1, 2, 3, 4, …

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 – p

There is a set of assumptions which, if valid, would lead to a binomial distribution. These are:

- A set of n experiments or trials are conducted.
- Each trial could result in either a success or a failure.
- The probability p of success is the same for all trials.
- The outcomes of different trials are independent.
- We are interested in the total number of successes in these n trials.

Under the above assumptions, let X be the total number of successes. Then, X is called a binomial random variable, and the probability distribution of X is called the binomial distribution.

Binomial Distribution Mean and Variance

- For a binomial distribution, the mean, variance and standard deviation for the given number of success are represented using the formulas

Mean, $\mu = np$

Variance, $\sigma^2 = npq$

Standard Deviation $\sigma = \sqrt{(npq)}$

Where p is the probability of success

q is the probability of failure, where $q = 1-p$

Applications

- Binomial distribution gives the possibility of a different set of outcomes. In real life, the concept is used for:

1. Finding the quantity of raw and used materials while making a product.
2. Taking a survey of positive and negative reviews from the public for any specific product or place.
3. By using the YES/ NO survey, we can check whether the number of persons views the particular channel.
4. To find the number of male and female employees in an organisation.
5. The number of votes collected by a candidate in an election is counted based on 0 or 1 probability.

Example

- Consider an exam that contains 10 multiple-choice questions with 4 possible choices for each question, only one of which is correct.
- Suppose a student is to select the answer for every question randomly. Let X be the number of questions the student answers correctly. Then, X has a binomial distribution with parameters n = 10 and p = 0.25. (Convince yourself that all assumptions for a binomial distribution are reasonable in this setting.)
- What is the probability for the student to get no answer correct? Answer:

$$P(X = 0) = 10! /(0!(10 - 0)!)\ (0.25)^0\ (1 - 0.25)^{10-0}$$
$$= (0.75)^{10}$$
$$= 0.0563$$

- What is the probability for the student to get two answers correct? Answer:

$$P(X = 2) = 10! /(\ 2!8!)\ (0.25)^2\ (1 - 0.25)^8$$
$$= 45 \cdot (0.25)^2 \cdot (0.75)^8$$
$$= 0.2816$$

- Binomial Mean and Variance. . .

It can be shown that

$$\mu = E(X) = np \text{ and}$$

$$\sigma^2 = V(X) = np(1 - p) .$$

For the previous example, we have

$$E(X) = 10 \cdot 0.25 = 2.5.$$

$$V(X) = 10 \cdot (0.25) \cdot (1 - 0.25) = 1.875.$$