Topic for the class:– Exploratory Data Analysis

Unit _2 : Title-Digital data – an Imprint

Unit2-syllabus

- **UNIT 2 Digital Data-An Imprint 9 hours, P - 2 hours** Type of data analytics (Descriptive, diagnostic, perspective, predictive, Prescriptive.) Exploratory Data Analysis (EDA), EDA-Quantitative Technique, EDA - Graphical Technique. Data Types for Plotting, Data Types and Plotting, Simple Line Plots, Simple Scatter Plots, Visualizing Errors, Density and Contour Plots, Histograms,
- Binnings, and Density, Customizing Plot Legends, Customizing Color bars, Multiple Subplots, Text and Annotation, Customizing Ticks.
- https://www.coursera.org/learn/data-visualization-r
-

Exploratory Data Analysis

- The purpose of exploratory data analysis (EDA) is to convert the available **data** from their raw form to an informative one, in which the main features of the data are illuminated
- We display and summarize data that are obtained from a sample.
- By means of this visualization and summaries, we try to explore the information hidden in the data
- These summaries are applied *only* to the data at hand; we do not attempt to make claims about the larger population from which the data is obtained

Exploratory Data Analysis

- When performing EDA, we should always:
  - –use **visual displays** (graphs or tables) plus **numerical summaries**
  - –describe the **overall pattern** and mention any **striking deviations** from that pattern
  - –**interpret** the results we got **in context**
- While exploring the Data, we should explore:
  - –Each variable individually (all Xs and Y)
  - –Each X with Y in pair
  - –Xs in pair
  - –All Xs and Y together

Exploratory Data Analysis

- The data that come from performing a particular measurement on all the subjects in a sample represent our observations for a single characteristic like country, age, education, etc.
- These measurements and categories represent a *sample distribution* of the variable, which in turn approximately represents the *population distribution* of the variable.
- One of the main goals of exploratory data analysis is to visualize and summarize the sample distribution, thereby allowing us to make tentative assumptions about the population distribution.

Key aspects of EDA include

- **Distribution of Data**: Examining the distribution of data points to understand their range, central tendencies (mean, median), and dispersion (variance, standard deviation).
- **Graphical Representations**: Utilizing charts such as histograms, box plots, scatter plots, and bar charts to visualize relationships within the data and distributions of variables.
- **Outlier Detection**: Identifying unusual values that deviate from other data points. Outliers can influence statistical analyses and might indicate data entry errors or unique cases.

Exploratory data analysis

- In statistics, **exploratory data analysis** (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
- Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.
- Exploratory data analysis is an analysis technique to analyze and investigate the data set and summarize the main characteristics of the dataset. Main advantage of EDA is providing the data visualization of data after conducting the analysis.

Univariate Analysis in EDA

- Univariate Analysis is an essential technique in Exploratory Data Analysis (EDA), allowing data scientists to understand individual variables in a dataset. It helps uncover patterns, anomalies, and insights that can drive decision-making.
- Univariate Analysis focuses on a single variable at a time, examining its distribution, summary statistics, and visual representation. It is a fundamental step before diving into more complex multivariate analyses.

Types of univariate Analysis

- **Central Tendency Analysis** 🎯 - *Mean*: Calculates the average value. *Example*: In a retail dataset, mean sales per month help identify sales trends.- *Median*: Identifies the middle value in a sorted dataset. *Example*: Median income provides a more robust measure of earnings, mitigating outliers' impact.- *Mode*: Finds the most frequently occurring value. *Example*: Mode identifies the most popular product in an inventory dataset.

Types of univariate analysis contd.

- **Dispersion Analysis** 💥 - *Range*: Measures the difference between the maximum and minimum values. *Example*: Range in temperature data can indicate climate variations.- V*ariance*: Shows the spread of data points from the mean.*Example*: Variance helps assess risk in financial investment portfolios.- *Standard Deviation*: Provides a more interpretable measure of data dispersion. *Example*: Standard deviation in test scores reveals how consistent students' performance is.

Types of Univariate Analysis contd.

- **Distribution Analysis** 📈 - *Histogram*: Displays the frequency distribution of data. *Example*: A histogram of employee ages can reveal age group - *Plot*: Highlights data distribution, skewness, and potential outliers. *Example*: Box plots help identify outliers in housing price data.- *Probability Density Function (PDF)*: Shows the likelihood of a variable taking specific values. *Example*: PDF of exam scores demonstrates grade distribution.

Types of univariate analysis contd.

- **Categorical Analysis** 📊 - *Bar Chart*: Represents the frequency of categorical data. *Example*: A bar chart can visualize the number of customers in different age groups.- *Pie Chart*: Illustrates the proportion of each category in a dataset. *Example*: A pie chart shows the market share of smartphone brands.

Significance of Univariate Analysis

- Univariate Analysis is the foundation of EDA, providing valuable insights into individual variables. By understanding central tendencies, dispersions, and distributions, data scientists can make informed decisions and draw meaningful conclusions from their data. Whether you're analyzing financial trends, healthcare data, or customer preferences, Univariate Analysis is your first step towards data-driven insights.

Key points in Univariate analysis:

1. **No Relationships:** Univariate analysis focuses solely on describing and summarizing the distribution of the single variable. It does not explore relationships between variables or attempt to identify causes.
2. **Descriptive Statistics:** Descriptive statistics, such as measures of central tendency (mean, median, mode) and measures of dispersion (range, standard deviation), are commonly used in the analysis of univariate data.
3. **Visualization:** Histograms, box plots, and other graphical representations are often used to visually represent the distribution of the single variable.

Summarizing the Data

- The data in general can be categorical or quantitative. For categorical data, a simple tabulation of the frequency of each category is the best non-graphical exploration for data analysis.
- For example, we can ask ourselves what is the proportion of high income

  professionals in our database:
- Given a quantitative variable, exploratory data analysis is a way to make preliminary assessments about the population distribution of the variable using the data of the observed samples.

Summarizing the data contd.

- The characteristics of the population distribution of a quantitative variable are its *mean*, *deviation*, *histograms*, *outliers*, etc.
- Our observed data represent just a finite set of samples of an often infinite number of possible samples.
- The characteristics of our randomly observed samples are interesting only to the degree that they represent the population of the data they came from.

Descriptive statistic

- Whenever we deal with some piece of data no matter whether it is small or stored in huge databases statistics is the key that helps us to analyze this data and provide insightful points to understand the whole data without going through each of the data pieces in the complete dataset at hand.

Descriptive statistics contd.

- In Descriptive statistics, we are describing our data with the help of various representative methods using charts, graphs, tables, excel files, etc. In descriptive statistics, we describe our data in some manner and present it in a meaningful way so that it can be easily understood. Most of the time it is performed on small data sets and this analysis helps us a lot to predict some future trends based on the current findings. Some measures that are used to describe a data set are measures of central tendency and measures of variability or dispersion.

Types of Descriptive Statistics

- Measures of Central Tendency
- Measure of Variability
- Measures of Frequency Distribution

Measures of central tendency

- Usually, frequency distribution and graphical representation are used to depict a set of raw data to attain meaningful conclusions from them. However, sometimes, these methods fail to convey a proper and clear picture of the data as expected. Therefore, some measures, also known as **Measures of Central Tendency** or **Average** are used as a single measurement to determine the main characteristics of the given series. Hence, the **Measure of Central Tendency** is a single value used for the representation of a complete set of data. Another name for Measure of Central Tendency is **Measure of Location.** Average is a typical value of the given data set to which most of the observations of the data fall closer than any other value. The three principal measures that are used in Statistical Analysis are **Arithmetic Mean, Median,** and **Mode.**

Mean

- One of the first measurements we use to have a look at the data is to obtain *sample statistics* from the data, such as the sample mean. Given a sample of *n* values, {*xi* }, *i* = 1, . . . , *n*, the *mean*, *μ*, is the sum of the values divided by the number of values, in other words:

$$\mu = \frac{1}{n}\sum^{n} x_i . \qquad (3.1)$$

Median

- The median is a centrally located value that splits the distribution into two equal portions, one including all values more than or equal to the median and the other containing all values less than or equal to it. The median is the ” middle ” element when the data set is organized in order of magnitude. As the median is established by the position of several values, it is unaffected if the size of the greatest value increases. The data or observations might be arranged in either ascending or descending order. In Statistics, Median is denoted by M.

Median

- **Step 1:** Firstly, arrange the given data in ascending or descending order.
- **Step 2:** Apply the following formula of Median:
  - $Median(M)=Size\ of\ [N+1]/2]th\ item Median(M)=Size\ of\ [2N+1]th\ item$
- Where, N is the Number of Items
- **Median in case of Odd and Even Number of Items**
- In case of **odd** number of items, Median is the Middle term of the observation. However, in case of **even** number of items, Median is the average of two middle terms and is determined by using the following formula:

Mode

- A given distribution can either be uni-modal, bi-modal, or multi-modal. Also, there can be a situation of no mode when each item of the series occurs an equal number of times.
- **No Modal Value:** When each observation of a series occurs the same number of times. **For example,** for the series 3, 4, 5, 4, 3, 5, 2, 2 there is no modal value as every item is occurring 2 times.
- **Uni-modal:** When one observation or item of a series occurs the maximum number of times. **For example,** the modal value or mode of the series 2, 4, 2, 5, 2, 6, 8 is 2 as it is occurring for the maximum number of times.
- **Bi-modal:** When two observations or items of a series have the same maximum frequency. **For example,** the mode of the series 4, 5, 4, 6, 5, 7, 8, 9 is 4 and 5 as both of these items are occurring two times.
- **Multi-modal:** When more than two observations or items of a series have the same maximum frequency. **For example,** the mode of the series 2, 3, 4, 4, 5, 5, 2, 1, 1, 7, 6, 9, 8 are 2, 4, 5, and 1.

Measures of dispersion

- Dispersion is the extent to which values in a distribution differ from the average of the distribution. Dispersion indicates a lack of uniformity in the size of items. It is the calculation of the extent to which numerical data is likely to vary about an average value. There are certain specific measures that help in determining the deviations from the central value, including Range, Quartile Deviations, Interquartile Deviations and Semi-Interquartile Deviations, Mean Deviations, Lorenz Curve, Standard Deviation, etc.

Dispersion

- Dispersion in statistics is a way to describe how spread out or scattered the data is around an average value. It helps to understand if the data points are close together or far apart.
- Dispersion shows the variability or consistency in a set of data. There are different measures of dispersion like range, variance, and standard deviation.

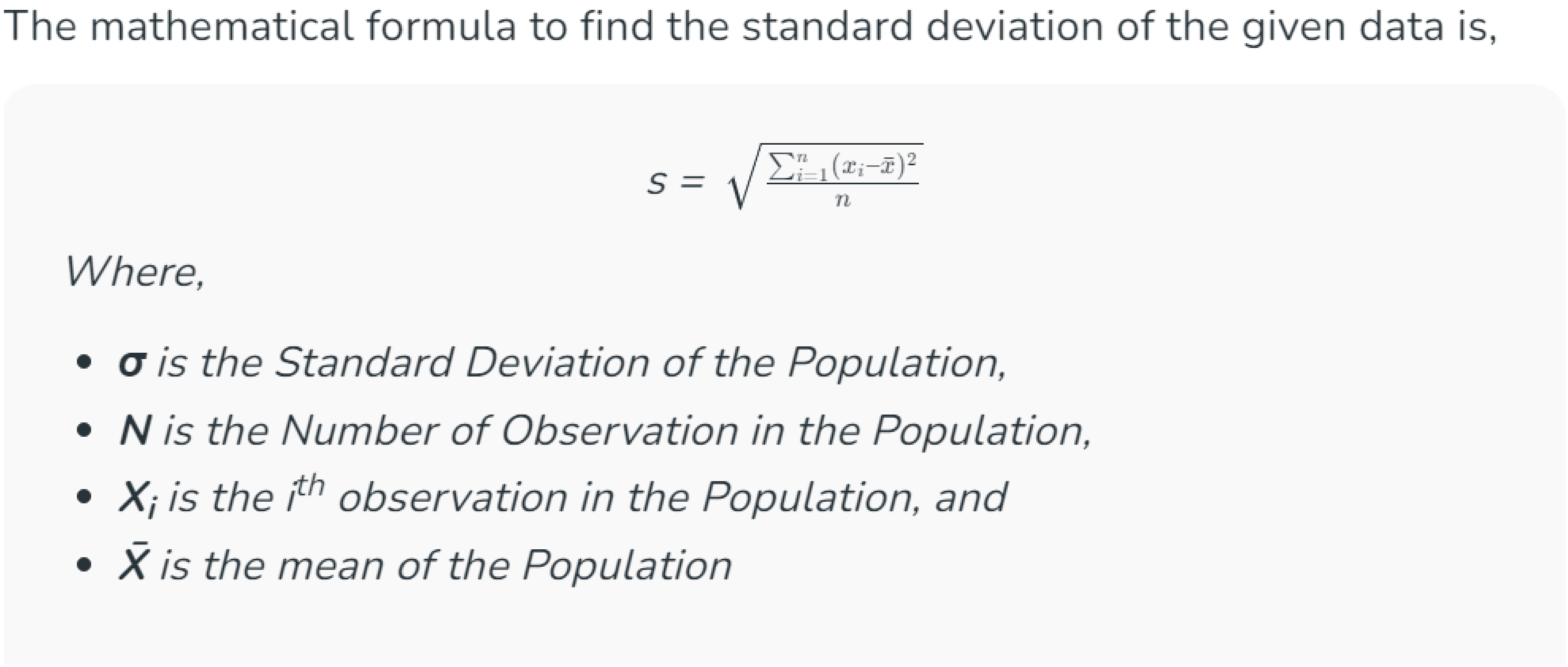
Variance and standard deviation

- Variance is the measure of how the data points vary according to the mean while standard deviation is the measure of the central tendency of the distribution of the data.
- The major difference between variance and standard deviation is in their units of measurement. Standard deviation is measured in a unit similar to the units of the mean of data, whereas the variance is measured in squared units.

Standard deviation

- How far our given set of data varies along with the mean of the data is measured in standard deviation. Thus, we define standard deviation as the “spread of the statistical data from the mean or average position”. We denote the standard deviation of the data using the symbol **σ**.
- We can also define the standard deviation as the square root of the variance.

Formula for population Standard deviation



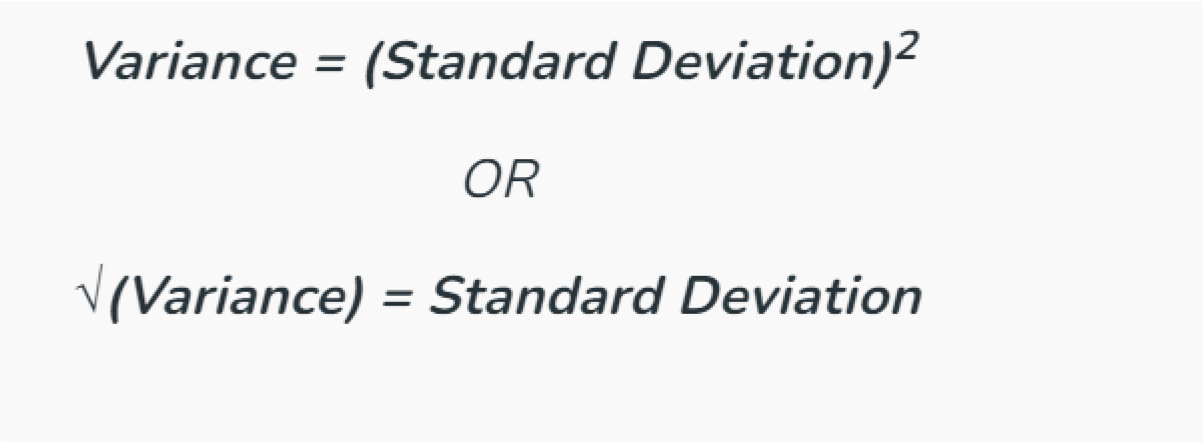
The mathematical formula to find the standard deviation of the given data is,

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

Where,

- ***σ*** *is the Standard Deviation of the Population,*
- ***N*** *is the Number of Observation in the Population,*
- ***X****$_i$* *is the ith observation in the Population, and*
- ***X̄*** *is the mean of the Population*

Relation between variance and standard deviation



Variance = (Standard Deviation)^2

OR

√(Variance) = Standard Deviation

THANK YOU